Master's thesis in Geography
Geoinformatics

TROPICAL ALTITUDINAL GRADIENT SOIL ORGANIC CARBON ESTIMATION WITH VIS-NIR (400-1000 nm) IMAGING SPECTROSCOPY

Niklas Sädekoski

2020

Supervisors:
Petri Pellikka
Mikko Toivonen

Master's Programme in Geography
Faculty of Science

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty	Osasto – Institution – Department
Faculty of Science	Department of Geosciences and Geography

Tekijä – Författare – Author

Niklas Sädekoski

Tutkielman otsikko – Avhandlings titel – Title of thesis

Tropical altitudinal gradient soil organic carbon estimation with VIS-NIR (400-1000 nm) imaging spectroscopy

Koulutusohjelma ja opintosuunta – Utbildningsprogram och studieinriktning – Programme and study track

Master's programme in geography, Geoinformatics

Tutkielman taso – Avhandlings nivå – Level of the thesis	Aika – Datum – Date	Sivumäärä – Sidoantal – Number of pages
Master's thesis, 30 credits	December 2020	63 + appendixes

Tiivistelmä – Referat – Abstract

Soil is the largest actively cycling terrestrial carbon pool, which has been severely distrubed in the last 100-200 years by human actions. To improve the situation, extensive monitoring of soil carbon and new methods for monitoring are required. This study demonstrates the capability of a portable hyperspectral device operating in the visible-near infrared (VIS-NIR) spectrum for soil organic carbon (SOC) prediction. Two multivariate methods, partial least squares regression (PLSR) and for this purpose previously untested lasso regression were used for prediction.

191 soil samples were collected from Taita Hills, Kenya. The samples represent a tropical altitudinal gradient with five land uses: agroforestry, field, forest, shrubland and sisal plantation. The samples were imaged with hyperspectral camera, Specim IQ in laboratory and in field conditions, and the carbon content of the samples was determined with a dry-oxidization analyzer. Three datasets were derived from the images, one containing the mean spectra of the complete imaged samples, one with segmented sub-image spectra and one with segmented sub-image spectra where outlier spectra were removed. Both multivariate methods were tested with all three datasets with good prediction accuracies (PLSR: R^2_{min} = 0.85, $RMSE_{min}$ = 0.78, lasso: R^2_{min}=0.85, $RMSE_{min}$=0.80), demonstrating the feasibility of both the device and lasso regression as SOC prediction tools. Using the segmented sub-image datasets improved the results with PLSR but had no significant effect on lasso regression prediction results.

While good results were gained with laboratory imagery, the field imaging conditions were difficult, and the data performed poorly. Future research should focus on finding solutions to reliably estimate SOC content in situ or with portable laboratory setups to make SOC measurements more widely accessible and agile for e.g. precision agriculture purposes.

Avainsanat – Nyckelord – Keywords

Soil organic carbon, VIS-NIR, imaging spectroscopy, partial least squares regression, lasso regression, Specim IQ

Säilytyspaikka – Förvaringställe – Where deposited

University of Helsinki electronic theses library E-thesis/HELDA

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty	Osasto – Institution – Department
Matemaattis-luonnontieteellinen tiedekunta	Geotieteiden ja maantieteen laitos

Tekijä – Författare – Author

Niklas Sädekoski

Tutkielman otsikko – Avhandlings titel – Title of thesis

Trooppisen korkeusgradientin maaperän hiilen arviointi kuvantavalla spektroskopialla näkyvän valon ja lähi-infrapunan alueella

Koulutusohjelma ja opintosuunta – Utbildningsprogram och studieinriktning – Programme and study track

Maantieteen koulutusohjelma, Geoinformatiikka

Tutkielman taso – Avhandlings nivå – Level of the thesis	Aika – Datum – Date	Sivumäärä – Sidoantal – Number of pages
Pro-gradu tutkielma	Joulukuu 2020	63 + liitteet

Tiivistelmä – Referat – Abstract

Maaperä on suurin aktiivisesti kiertävä maanpäällinen hiilivarasto, joka on heikentynyt suuresti viimeisen 100-200 vuoden aikana ihmistoiminnan seurauksena. Tilanteen parantamiseksi vaaditaan laajamittaista maaperän hiilen seurantaa ja kehittyneempiä metodeja tätä varten. Tässä tutkimuksessa demonstroidaan näkyvän valon ja infrapunan aallonpituuksilla toimivan hyperspektrikameran toimivuutta maaperän orgaanisen hiilen ennustamisessa. Tähän käytetään kahta monimuuttujamenetelmää, PLS-regressiota, sekä lasso regressiota, jota ei ole aikaisemmin tähän tarkoitukseen käytetty.

191 maaperänäytettä kerättiin Taitavuorilta Keniasta trooppiselta seudulta nousevan rinteen ympäriltä, viiden eri maankäytön alueelta, jotka ovat: peltometsäviljely, pelto, metsä, pensasmaa sekä sisal plantaasi. Näytteet kuvattiin hyperspektrikamera Specim IQ:lla sekä laboratoriossa, että kentällä. Kuvista tuotettiin kolme datasettiä, yksi kuvien keskiarvoisella spektrillä, toinen segmentoitujen kuvien osien keskiarvoisilla spektreillä ja kolmas segmentoitujen kuvien osien keskiarvoisilla spektreillä siten, että ääriarvot suodatettiin pois. Sekä PLS-regressio- sekä lasso regressiomallit antoivat hyviä tuloksia kaikilla dataseteillä (PLSR: R2min = 0.85, RMSEmin = 0.78, lasso: R2min=0.85, RMSEmin=0.80) viitaten sekä laitteen tuottaman datan, että lasso regression soveltuvan maaperän orgaanisen hiilen mallintamiseen. Segmentoitujen osa-kuvien käyttö mallien opettamisessa paransi tuloksia PLSR malleissa, mutta ei vaikuttanut merkittävästi lasso regressiomallien tuloksiin.

Vaikka laboratoriossa kuvannettu data antoikin hyviä tuloksia, kenttäolosuhteissa kuvaaminen oli haasteellista ja tulokset tällä datalla olivat heikkoja. Tulevien tutkimusten tulisikin keskittyä kenttämenetelmien kehittämiseen ja löytämään ratkaisuja maaperän hiilen luotettavaan mittaamiseen suoraan maasta, tai lähellä tutkittavaa kohdetta siirreltävien laboratoriojärjestelyiden avulla. Tämä parantaisi hiilimittausten saavutettavuutta ja mahdollistaisi niiden paremman hyödyntämisen esimerkiksi täsmäviljelyssä.

Avainsanat – Nyckelord – Keywords

Maaperän hiili, spektroskopia,PLS-regressio, lasso regressio, Specim IQ

Säilytyspaikka – Förvaringställe – Where deposited

University of Helsinki electronic theses library E-thesis/HELDA

Muita tietoja – Övriga uppgifter – Additional information

Table of contents

1 Introduction

Soil is the largest actively cycling terrestrial carbon pool with approximately 1500-2000 Pg of carbon in the depth of 1 m globally (Janzen 2004). This pool has been severely disturbed mainly by the expansion of agriculture in the last 100-200 years by removing the above-ground vegetation as input sources of carbon to soil (Houghton 1999). In tropical Africa a land conversion from natural to agricultural can cause a 75% depletion to the SOC pool according to Lal (2004). In addition to contributing to the current climate change by adding carbon to the atmosphere, the loss of carbon weakens the food productivity, nutrient and water retention, and the overall structure of the soil (Ontl & Schulte 2012, FAO and ITPS 2018). This problem is already well recognized and countering actions have been proposed in form of soil carbon sequestration, which means transferring atmospheric CO_2 into soil as long-living pools by means such as crop management practices, afforestation and agroforestry (Lal 2004).

To actively improve the situation and to find the best places to implement these practices, current information and extensive monitoring of soil carbon content is required. This has proven to be challenging especially on large scale and estimates of the size of the soil carbon pool have large variance (Todd-Brown et al. 2013). Traditional method for SOC mapping has been collection and laboratory analysis of soil samples, which is time consuming and expensive regardless of the scale, since the number of the samples should be high to attain reliable results and great deal of manual labor is required to process them (Peón et al. 2017). Most large-scale maps are already produced with remotely sensed spectral data acquired by airborne or satellite sensors, but these are still very approximate and suffer from inaccuracies caused by atmospheric absorption, illumination variations and low signal-to-noise ratio (Peón et al. 2017) as well as physical obstructions such as vegetation or manmade structures (Jensen 2009). Spectroscopy has also been proposed as a nondestructive and fast substitute for soil carbon analyzes in laboratory and field with laboratory spectrometers and portable imaging spectrometers, and laboratory spectrometers have already proven to be about as precise as traditional methods (Doetterl

et al. 2013, Aldana-Jague et al. 2016b). Although widely researched, consensus among scientists is that the potential of spectroscopy in the field of soil studies and SOC measuring is still to be exploited (Ben-Dor et al. 2009).

This study aims to deepen the field by experimenting the applicability of a previously untested portable hyperspectral camera, Specim IQ (Specim, Spectral Imaging Oy Ltd, Finland) for estimating SOC. IQ is a handheld device with a weight of 1.3 kg, with spectral range of 400-1000 nm, 7 nm spectral resolution and 204 bands. In addition to capturing raw spectral intensity data, it's able to produce a reflectance transformation on the fly reducing the required post processing steps (Specim IQ manual.). It has been previously used at least for plant phenotyping and disease detection (Behmann et al. 2018). Due to its small size and relatively easy usability, and low requirement for sample preparations, it could be utilized for determination of SOC close within the study areas without a need to transport the samples to be analyzed in laboratory. Applying an imaging spectrometer, it is also possible to process multiple samples simultaneously, as presented by O'Rourke et al. (2011). The limitation of IQ is its spectral range, as soil studies would require middle infrared wavelength, e.g. Rossel et al. (2006) applied near infrared area up to middle infrared until 2500 nm for SOC assessment (Rossel et al. 2006). However, moderate results have been obtained with limited wavelengths earlier (Jung et al. 2015) and many bandwidths correlating with SOC have been found also in the VIS-NIR region (Yang & Li 2013, Ben-Dor et al. 1999, Rossel et al. 2006).

The samples for this study were collected from Taita Hills, Kenya (Wundanyi 03°23'54"S 38°21'37"E), a mountainous area with diverse land cover from savannah shrubland at 800 m altitude (above sea level) to indigenous mountain forests at 2200 meters (Pellikka, PKE et al. 2018). Intensive agriculture is practiced in the highlands, while the lowlands are used for livestock management, dryland farming, sisal production and conservation. Agricultural expansion has led to loss of forests and bushlands (Pellikka, Petri K. E. et al. 2009, Pellikka, PKE et al. 2018), thus evidently to decreased SOC. Therefore, the area could benefit of new kind of soil monitoring techniques. The land cover and land use changes and their consequences to ecosystem services and climate change have been studied by the

University of Helsinki since 1989 (Pellikka, P. 1990). The hyperspectral data from IQ could also be used as a reference/calibration data to remote sensing data collected by satellites and airborne hyperspectral remote sensing imagery collected by AisaEAGLE sensor (Specim, Spectral Imaging Oy Ltd, Finland) (Heiskanen et al. 2019), which has the same spectral range than IQ.

Partial least squares regression (PLSR) has been the most common method of estimating SOC from spectral data (Vasques et al. 2008). PLS is closely related to principal components regression (PCR), where the dimensionality of collinear variables is reduced by finding lower number of new latent variables that can be used to model the dependent variable under study (Martens & Naes 1992). PLSR operates similarly, but unlike PCR, PLSR also uses the dependent variable(s) while finding the new variables, usually ending in better relation between latent variables and the explained variable (Martens & Naes 1992). In addition to using PLSR, in this study the applicability of lasso regression for SOC prediction is tested. Lasso is a regularization technique for multiple linear regression (MLR), that is used to lower the amount of variables by adding a penalty term to the MLR coefficient estimation function (Tibshirani 1996), and is slightly simpler and more intuitive in relation to PLSR. Models and analysis in this work were mostly implemented using open source Python software and libraries.

The aim of this study is to assess the feasibility of handheld imaging spectroscopy applying Specim IQ for estimating SOC content in field conditions in tropical Africa and in a laboratory, and to test the potentiality of lasso regression as a multivariate method for SOC estimation from hyperspectral data.

2 Background

2.1 Soil organic carbon (SOC)

Soil organic carbon is a product of partial decomposition of any living organisms in the soil (FAO and ITPS 2018), plant roots being the main input source (Ontl & Schulte 2012). Carbon is usually found in the soil as carbon-based compounds such as carbohydrates and proteins, only rarely as pure carbon from burnt material (Hall 2008). About third of global SOC stocks are in forests, third in grasslands and savannas and the rest in wetlands, croplands and other biomes (Janzen 2004).

Highest SOC levels can be found from the cold and wet northern latitudes, where decomposition is slower than the rate of photosynthesis. Warm and wet weather in tropics lead to high primary productivity and fast decomposition resulting in medium levels of SOC, whereas arid areas with low primary productivity have low levels of SOC (Ontl & Schulte 2012, Scharlemann et al. 2014). In a local scale, there are many factors affecting the levels of SOC, including soil texture, minerology, and erosion and deposition processes (Ontl & Schulte 2012), as well as topography (Cardinael et al. 2017) and soil type (Zhao et al. 2006).

SOC should not be mixed with soil organic matter (SOM), but is a part of it. SOM is a larger concept that includes all organic material in the soil from fresh plant residues to highly decomposed humus and the amount of SOC is thus naturally highly correlated with the amount of SOM (Ontl & Schulte 2012).

Soil organic matter is a key component of a healthy soil as it contains nutrients and increases the retention of water, improves the overall structure of the soil and reduces erosion. High SOM and thus SOC levels therefore result in better food productivity of the soil by providing nutrients and water (FAO and ITPS 2018, Ontl & Schulte 2012).

2.1.1 Global carbon pools and the carbon circle

With approximately 1500-2000 Pg of carbon in various forms in the depth of 1 m, soil is the largest actively cycling terrestrial pool of carbon, compared to about 785 Pg C in the atmosphere as CO_2 and 400-600 Pg carbon stocks in biota of which most is in the forests

(Janzen 2004, Scharlemann et al. 2014). The largest carbon reservoir is still in the oceans with about 39 000 Pg C, mainly in the deep seabed layers and not actively circulating (Janzen 2004).

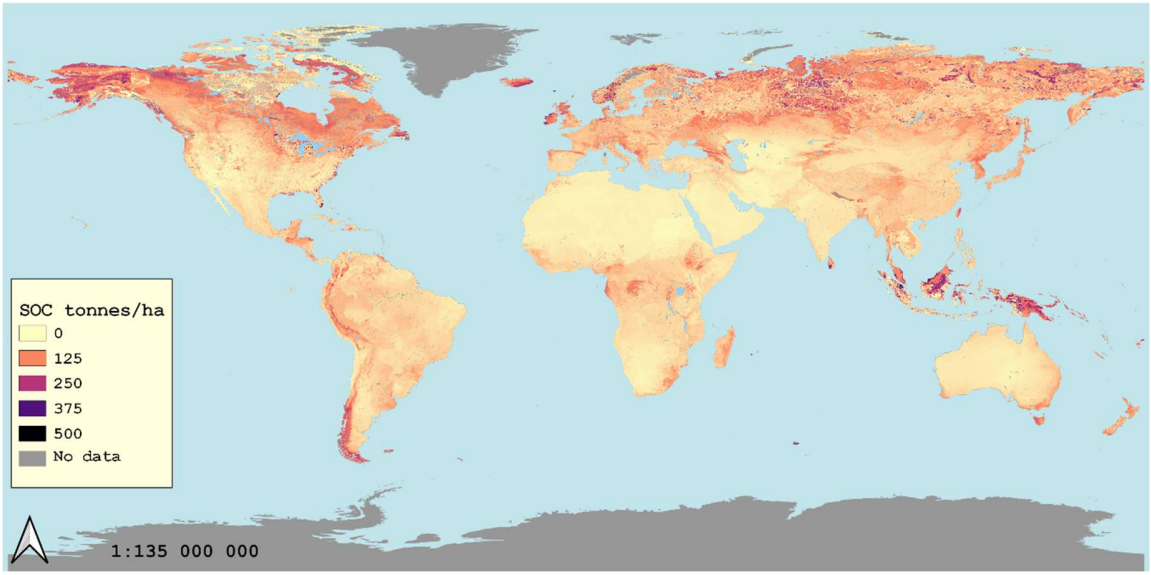

Figure 1. Global SOC distribution in 30 cm depth (FAO and ITPS).

Most of the global SOC is stored at the northern latitudes where slow decomposing conditions allow carbon to accumulate (Scharlemann et al. 2014), with largest reservoirs in the permafrost regions of Russia and Canada (Figure 1. Global SOC distribution in 30 cm depth (FAO and ITPS)Figure 1) (FAO and ITPS 2018).

All carbon pools are connected, and carbon circulates between them in various ways. Vegetation is the gateway between atmospheric CO_2 and terrestrial biomass via photosynthesis, which binds about 120 Pg C from the atmosphere to vegetation per year, although about 50 % of that is released back very soon by plant respiration (Janzen 2004). From vegetation, carbon is transferred to fauna by consumption and to soil by decomposition of living organisms (FAO and ITPS 2018, Janzen 2004). Meanwhile, fire and heterotrophic respiration by soil microorganisms return approximately 60 Pg C per year back to atmosphere as CO_2 thus closing the circle (Janzen 2004). In oceans carbon circulates mainly between the atmosphere and the surface water whereas the deep ocean reservoirs

are more stable and are affected for example by several water-mixing effects and sinking of dead organisms and fecal pellets (Post et al. 1990).

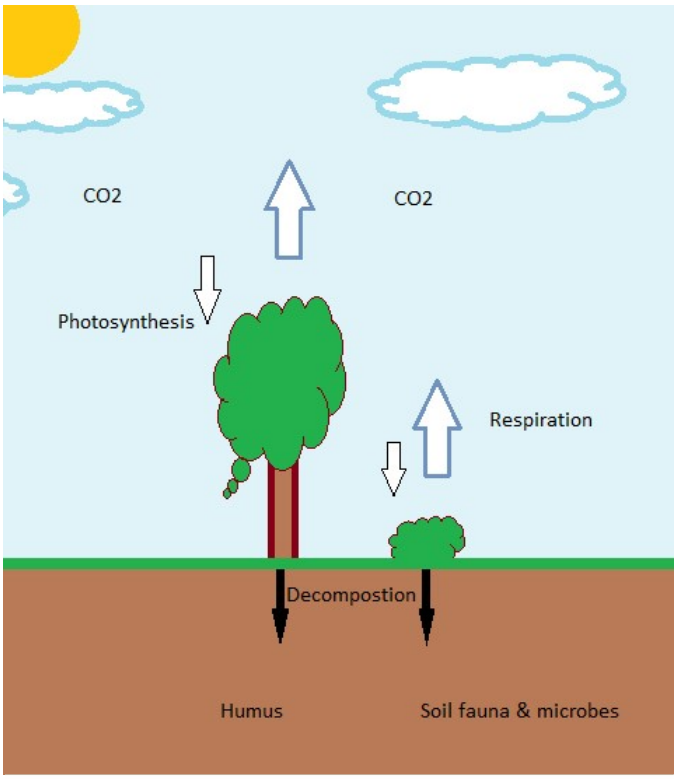

Figure 2 Carbon cycle between atmosphere and soil.

2.1.2 Human influence

The exponential growth of human population has been impacting the global carbon cycle by increasing the amount of CO_2 in the atmosphere since the industrial revolution (Ontl & Schulte 2012). Carbon inputs to atmosphere are caused mainly by burning of fossil fuels and changes in land use (Janzen 2004), with about two thirds caused by the former and one third by the latter (Lal 2004). According to Lal (2004), land use conversion from natural vegetation to agricultural landscapes in temperate regions can cause a depletion of SOC pool by 60%, and over 75% in the tropics. Conversion in arid and semi-arid regions from natural land use to croplands can, however conversely increase the amount of SOC (Ramesh et al. 2019).

Thus, the expansion of agriculture has had the highest impact on SOC levels with worldwide cultivated areas quadrupled between years 1850-1990, reducing forests by 17%, mostly in the tropics (Houghton 1999). Agriculture reduces and redirects the carbon stocks not only by removing the original biomass (e.g. forests) binding the carbon but also by constantly taking away the C absorbed into the crops by photosynthesis as yield, as illustrated by Janzen (2004). This combined with the even severe problem of burning fossil fuels in the last century has led to the imbalance in the global carbon cycle where rising CO_2 levels in the atmosphere are causing warming temperatures and climate change all around the globe (Ontl & Schulte 2012).

2.1.3 Carbon sequestration

There are various means to improve SOC levels, such as better crop management practices, afforestation, agroforestry and conservation planning, to name a few (Janzen 2004, Lal 2004, Scharlemann et al. 2014, Ramesh et al. 2019). These means are collectively called soil carbon sequestration, which implies transferring atmospheric CO_2 into soil/biomass as a long-living pools that are not immediately returned (Lal 2004). As such, soil sequestration is a natural process, but as previously illustrated, humans have distorted the rate of carbon stored/carbon released and hence reviving actions are needed. According to Lal (2004), the potential soil carbon sink capacity of managed ecosystems is close to the cumulative historic losses at 55 – 78 Gt. In addition to the potential capacity, soil carbon sequestration has been seen as a good option because of its cost effectiveness and its positive effects on soil along with the stored carbon, e.g. better soil fertility and biodiversity (Janzen 2004). Most management practices are directed to agriculture, since it has had the greatest impact on SOC losses but also because improving SOC levels is profitable for the soils under cultivation. The practices can be divided to preserving actions that reduce the loss of C in soil and adding actions that increase the amount of C by inputs. Reducing practices include reduced tillage or no tillage, erosion control, agroforestry and diverse cropping systems, while adding actions can be e.g. the input of manure, compost and other fertilizers or management of crop residues (Ontl & Schulte 2012, Lal 2004, Scharlemann et al. 2014).

2.1.4 Challenges

Even though in many cases these management practices offer a win-win resolution in terms of storing C and soil productivity/overall soil well-being, there are still many challenges and uncertainties linked to soil carbon sequestration.

First of all, monitoring the stocks is proven to be quite challenging. Global SOC estimates have large variation between studies, e.g. Todd-Brown et al. (2013) compared 11 models with a range of 510-3040 Pg C for global soil carbon stock. As there are numerous factors affecting the carbon stabilization mechanisms in soil, from natural conditions like temperature, rainfall and topography to many human activities from fertilizing to land use modifications, it is still not well understood how these all function together (Ramesh et al. 2019, Janzen 2004). Janzen (2004) adds, that it's difficult to show what exactly is then reason behind the changes in SOC levels when they occur, and questions our ability to measure the gains/losses in short temporal scale (e.g. a year).

The potential of soil C sequestration is also limited and can be achieved in 20 to 50 years and thus cannot be seen as long term solution to rising CO_2 levels (Lal 2004, Janzen 2004). The use of site suitable practices is crucial to achieve the best results and to avoid unwanted setbacks like acidification (Scharlemann et al. 2014) or increased CO_2 emissions (Ramesh et al. 2019). Global hotspots where actions are most needed are located in the tropical regions, which account for almost all of the carbon losses of recent decades (Houghton 1999). These regions are also in many cases lacking in institutions, infrastructure and resources to focus on soil sequestering as an issue (Lal 2004).

All these combined with the uncertain effect of the ongoing climate change and population growth to the global carbon cycle and SOC stocks and fluxes create a need to keep studying and improving our models and estimates on the subject (Scharlemann et al. 2014, Janzen 2004).

2.2 Remote sensing and imaging spectroscopy

Remote sensing (RS) is defined as "the measurement or acquisition of information of some property of an object or phenomenon, by a recording device that is not in physical or

intimate contact with the object or phenomenon under study" by the American Society for Photogrammetry and Remote Sensing (Jensen 2009). According to this definition, the distance of "remote" is not defined and thus close-up images taken with mobile phone can be interpreted as remote sensing data as well as satellite images. Commonly RS however refers to image data collected with a sensor mounted on e.g. Unmanned Aerial Vehicle (UAV), aircraft or satellite, as defined by Richards (1999).

2.2.1 Principles of remote sensing

Regardless of the scale the principle is the same, where most of the remote sensing sensors record electromagnetic radiation (EMR) emitted by or reflected from the subject of study (Jensen 2009). More precisely, RS sensors measure *radiance* (L_λ), which is the radiant intensity per unit of projected source area in a specified direction, measured as watts per meter squared per steradian (W m^{-2} sr^{-1}). (Jensen 2009)

Remotely sensed image is constructed with sensor detectors that are electrically charged by the incoming light by an amount directly related to the incident radiant energy. These charges are converted to digital brightness values representing the radiance received by the sensor from the instantaneous field of view (IFOV), that we can observe as an image. Most common digital sensors are charge-coupled device (CCD) and complementary metal oxide semiconductor (CMOS). (Jensen 2009).

2.2.1.1 Push broom and Whisk broom scanners

Two most common sensor system types at least in satellites and aircrafts are push broom and whisk broom scanners (L3Harris Geospatial. 2020). The names refer to the way the data is collected.

In whisk broom (or across track) scanners, a mirror rotating perpendicular to the moving direction scans the target area one pixel at a time and reflects the incoming light to just one detector (thus "whiskbrooming" the current row of pixels) (Figure 3)(Jensen 1996).

Push broom (or along track) sensors have a dedicated detector for each pixel along track, forming a linear array of detectors. (Jensen 1996). This allows the simultaneous scanning of the whole row of pixels ("pushbrooming" the image) and thus longer exposure time for each pixel to get a stronger signal of the incoming radiance (Figure 3)(L3Harris Geospatial. 2020).

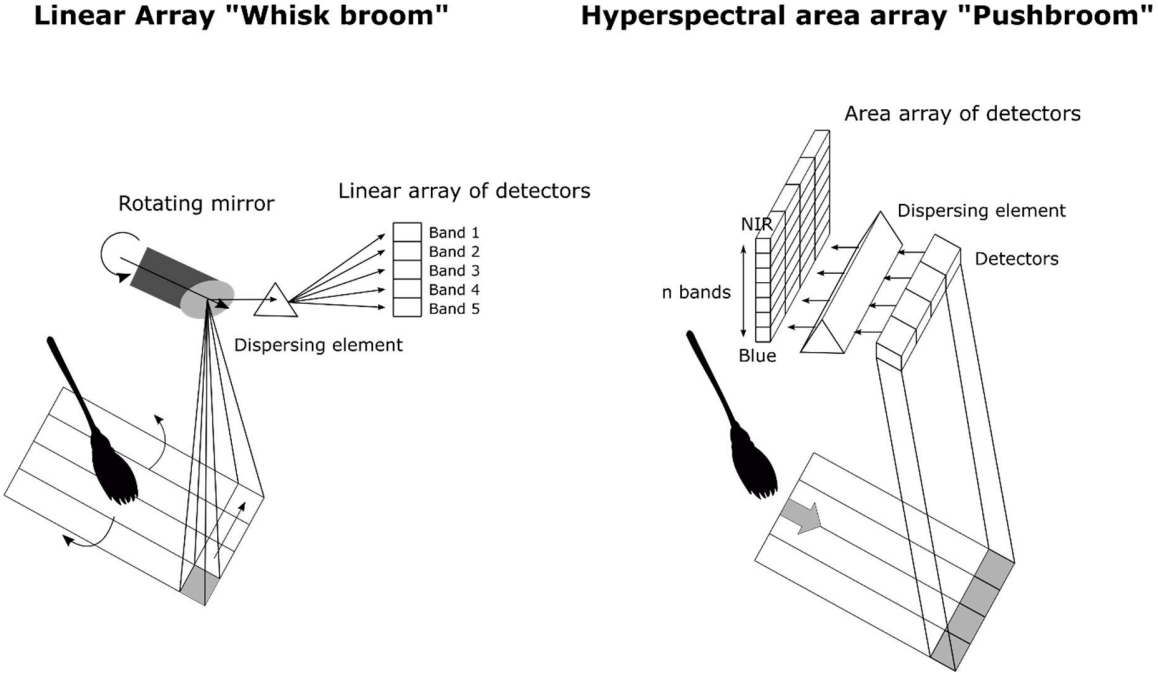

Figure 3 Whiskbroom and pushbroom scanners. Hyperspectral systems use spectrometers or prisms to further disperse the incoming light to linear or area arrays with as many elements as there are spectral bands in the final product (figure modified from (Jensen 1996)).

2.2.1.2 Passive and active sensors

Sensors are also divided to passive sensors (most satellites, cameras, video recorders etc.), that record the reflected or emitted EMR produced by some external source, usually the sun, and active sensors which transmit artificial EMR and record the backscatter of the sent energy after it has interacted with the terrain under study (Jensen 2009). Examples of active sensing methods are Radio Detection and Ranging (RADAR, transmits and receives microwave energy), Light Detection and Ranging (LiDAR, transmits and measures laser pulses) and Sound Navigation Ranging (SONAR, emits sound waves and records echoes).

2.2.1.3 Spectral range and resolution

RS sensors differ also in their ability to differentiate the electromagnetic spectrum. Spectral range or wavelength range is the range where the sensor operates on the spectrum, e.g. normal cameras usually use the visible light from 400 to 700 nm to form a picture with three different bands (red, green and blue). Most modern remote sensing sensors are *multispectral*, which means that they are able to record energy in multiple more specific bands of the spectrum from visible light to microwave (Richards & Richards 1999).

Spectral resolution refers to the number and width of the bands that the RS instrument is sensitive to. For example, the Landsat 7 Enhanced Thematic Mapper satellite sensor system records data in 8 bands with a spectral resolution of 150 nm in band 4 (from 750 – 900 nm), meaning the band 4 is sensitive to EMR in the range of 750 -900 nm. (Jensen 2009).

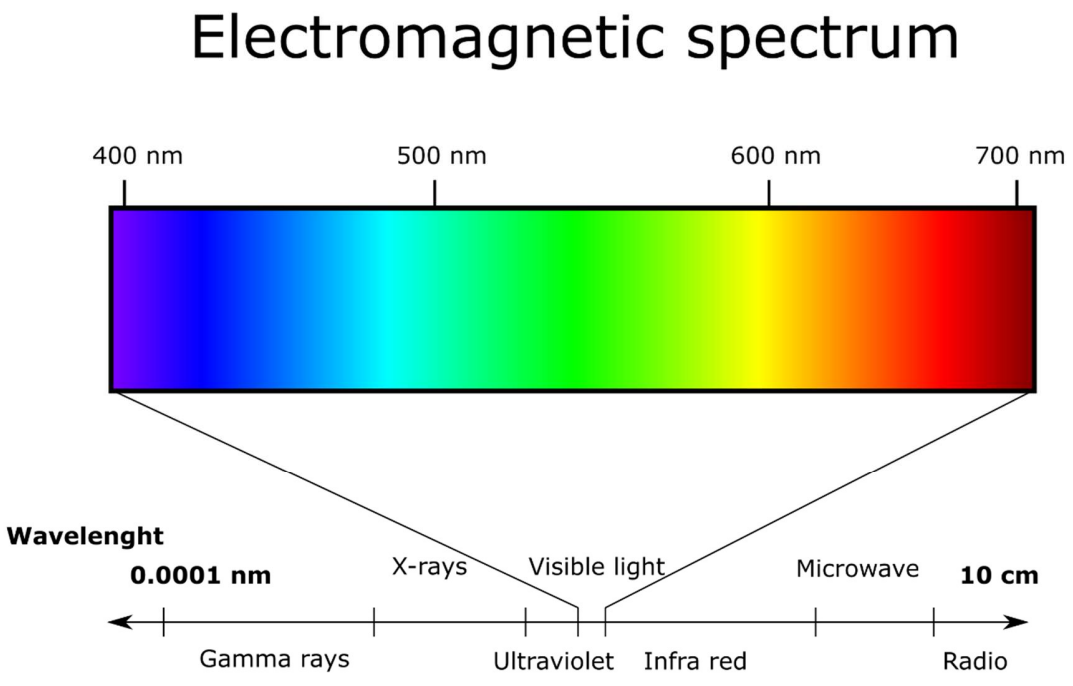

Figure 4. The electromagnetic spectrum.

2.2.2 Imaging spectrometers

Hyperspectral sensors, also called as imaging spectrometers, can record hundreds of bands with a very narrow spectral resolution, for example 7 nm as in Specim IQ, producing a complete spectra for each element (smallest separable object) of the acquired image, in the spectral range of the sensor (Richards & Richards 1999) (Figure 5). This can be utilized to study different materials very precisely as many materials and chemicals have very small spectral absorption windows of 20 – 40 nm, which can be used to identify these materials and also provide information about the state of the material (Jensen 2009). Before analyzations, the at-sensor radiance is usually transformed to *reflectance,* which is the ratio of incident radiance to the observed target and the reflected radiance captured by the sensor. Reflectance spectrums are comparable regardless of the effects of current lighting, absorption and scattering effects, but require data of the current conditions to be gathered along the imaging. (Jensen 2009)

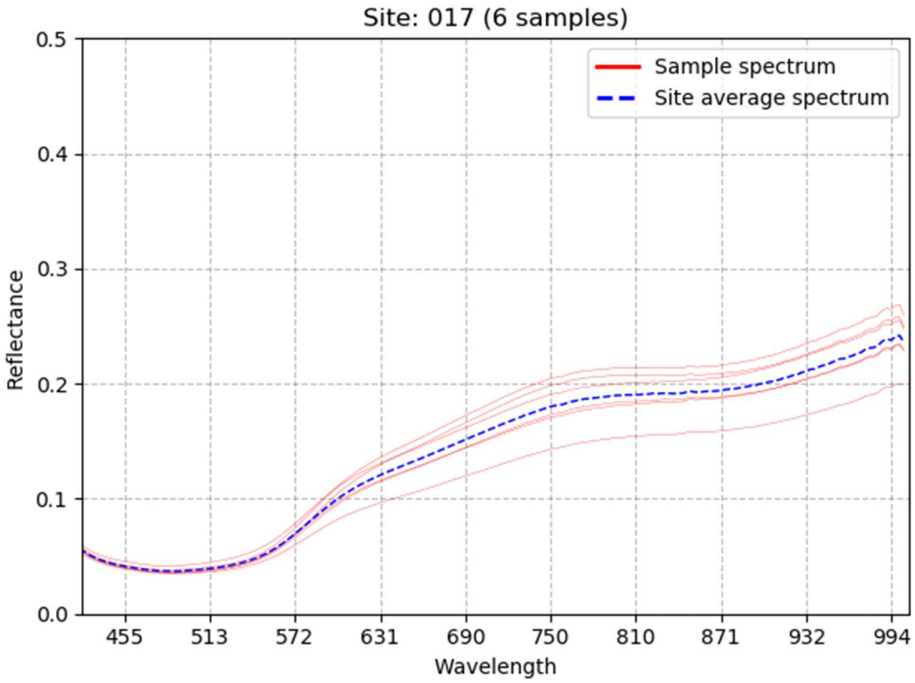

Figure 5. Example percentage reflectance curves of agricultural soils from Taita hills, Kenya, acquired with Specim IQ handheld hyperspectral camera.

Most imaging spectrometers are passive pushbroom or whiskbroom sensors, where the radiant flux entering the sensor is passed to a spectrometer, which disperses it onto a linear or area array of detectors (Jensen 2009) (Figure 3). Each detector in the array is dedicated to a specific spectral band and thus each image element is saved in as many spectral bands as there are detector elements in the sensor (Jensen 2009). In pushbroom sensors, there are as many linear arrays of detectors as there are pixels in the image.

Hyperspectral images are commonly called as *data cubes* as in addition of the two traditional dimensions of 2D image, the spectral bands form a third dimension (Figure 6).

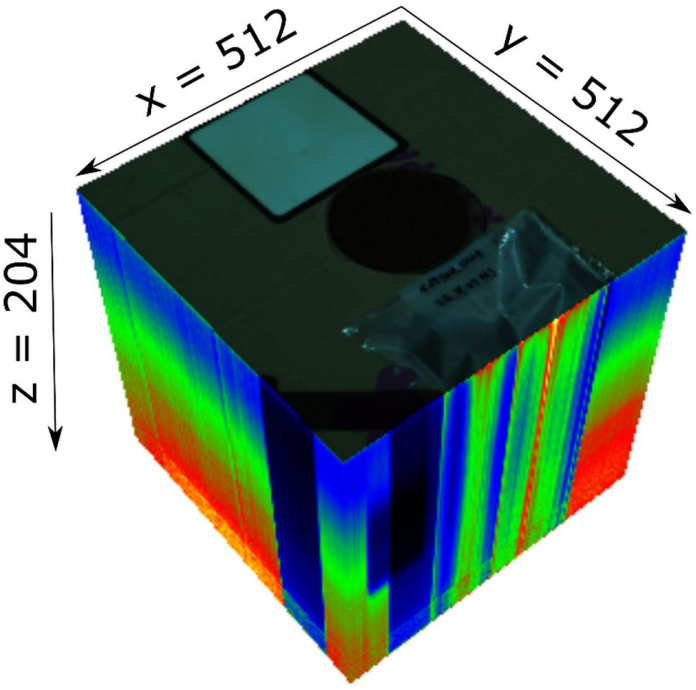

Figure 6. Specim IQ data cube with 512 x 512 rows and lines and 204 spectral bands.

Downside of three-dimensional hyperspectral data is its size, as one data cube comprises of as many grayscale images as there are spectral bands. This sets a requirement for large amount of storage space and also powerful computers to effectively process the data. Especially satellite and airborne sensors also suffer from various radiometric distortions in the data caused by atmospheric effects and instrumentation errors (Richards & Richards 1999)

2.3 Measuring SOC

Large scale measuring of SOC stocks has been and is still challenging (Scharlemann et al. 2014). Traditional methods for mapping SOC has been the collection and analysis of soil samples by carbon oxidization methods, which is time consuming and expensive since the number of samples needed is very high and the analyses require a lot of manual labor (Peón et al. 2017). Remote sensing and spectroscopy has been recognized as a substitute method especially for large scale mapping of topsoil SOC, but also for soil sample analyses with spectrometers (Scharlemann et al. 2014).

2.3.1 Traditional methods

Two most commonly used traditional methods for analyzing SOC content from soil samples are wet combustion and dry combustion.

In wet combustion the samples are heated with a potassium dichromate-sulfuric acid-phosphoric acid solution ($K_2Cr_2O_7$-H_2SO_4-H_3PO_4) creating a temperature of 210°C to oxidize the organic matter into CO_2 (Chatterjee et al. 2009). There are several modifications of wet combustion, with Walkley & Black (1934) being widely used due to its simplicity, rapidness and minimal requirement for equipment (Nelson & Sommers 1996). Wet combustion methods however do not usually manage to oxidize all present C from the sample, but the results need to be corrected by some factor, depending on variables such as soil type (Chatterjee et al. 2009). Thus, wet combustion methods are not very exact and the results should be considered as approximate (Nelson & Sommers 1996).

Dry combustion methods use high temperatures to oxidize the soil carbon and measure the generated CO_2 by either calculating the mass loss-on-ignition (LOI) or with automated analyzers able to measure the amount of CO_2 (Chatterjee et al. 2009). LOI methods use temperatures of 500-550°C and require several hours of heating of oven-dried samples in a muffle furnace (Heiri et al. 2001). Automated dry oxidization analyzers, which are considered as the current standard method for SOC content determination, use very high temperatures (950-1150°C) to oxidize the samples. CO_2 is then separated from other gases and measured by thermal conductivity, mass spectrometry or infrared gas analyzing

methods (Smith & Tabatabai 2003). This is the most rapid, precise and simple method to determine SOC content, but bear a downside as the analyzer devices are very expensive and thus not available for extensive use (Chatterjee et al. 2009).

2.3.2 Soil spectroscopy

Although not without difficulties, remote sensing and spectroscopy has already proven its potential as a substitute nondestructive method for SOC mapping. In laboratory conditions it comes close to be as precise as traditional laboratory analysis methods (Doetterl et al. 2013, Aldana-Jague et al. 2016a). At the World Agroforestry center (ICRAF) in Nairobi, Kenya, spectroscopy is used as the main method for soil analysis (ICRAF. 2020) . Laboratory based hyperspectral imaging was also found to be the cheapest and most rapid alternative in comparison to wet chemistry and dry combustion by O'Rourke et al. (2011).

Most studies have still focused on testing new methods and finding the best spectral response bands for SOC and other soil properties with laboratory spectrometers (Ben-Dor et al. 1997, Bartholomeus et al. 2008, Peng et al. 2014). Portable imaging spectrometers, which potentially offer better spatial coverage and better reference for airborne sensors, have also been tested as a substitute to traditional point spectrometers in laboratory and in field with promising results (Jung et al. 2015, Steffens & Buddenbaum 2013). Controlled laboratory-based measurements with high level spectrometers have laid important understanding of the spectral principles of soil and its constituents that can be used in developing larger scale spectral applications in addition to offering rapid and relatively inexpensive laboratory analysis method (Ben-Dor et al. 1999). Spectrometers have also been used in field conditions with moderate results by e.g. Bayer et al. (2016) with a maximum R^2 value of 0.62 for the field data predictions.

In-field research and larger scale remote sensing mapping outside laboratory with airborne and satellite imaging hyperspectral sensors still has several limitations e.g. atmospheric absorption, illumination variations and the low signal-to-noise ratio of the sensors (Peón et al. 2017). Also soil reflectance is highly influenced by the soil surface characteristics like

moisture and soil roughness, and the remote sensed data needs to be calibrated to the current conditions and usually still requires soil samples from the area (Aldana-Jague et al. 2016a). Vegetation and other physical obstructions further complicate or completely prevent the use of remote sensing for studying soil properties where the soil is partly or completely covered (Jensen 2009). Upon these issues are taken into account by image processing and calibrating the data with proper reference data, remote sensing has been proven to be a relatively cheap and fast substitute for developing large scale SOC maps (Peón et al. 2017). Best areas to utilize airborne remote sensing for SOC monitoring have been found in agricultural lands where the soil is uncovered, and monitoring is needed. Stevens et al. (2010) employed aerial hyperspectral imagery captured from airplane to map SOC in Luxemburg croplands with promising results, although noting that the models were mostly regional. Hbirkou et al. (2012) and Bayer et al. (2016) both used HyMap sensor in their SOC studies with slightly varying results, the R^2 values ranging from 0.83 to 0.62).

Although lot of research has been done, consensus among scientists is that the potential of imaging spectroscopy (IS) for soil studies has not been reached (Ben-Dor et al. 2009). The most promising applications of IS would be in the field for mapping the spatial variability and changes of SOC, as well as for precision agriculture (Morgan et al. 2009), but this requires more research on new methods and improved models to reach the possible potential.

2.3.2.1 Soil spectral properties

The spectral reflectance of soils is affected by several factors, which include in addition to the carbon content, texture, moisture content, iron-oxide content, salinity and surface roughness (Jensen 2009). This sometimes complicates the spectral analysis of soil, because some of these factors have very similar effects. For instance, the general relation between SOC and reflectance in the visible spectrum is that reflectance decreases with more organic content (Ben-Dor et al. 1997). Moisture content has the same effect, where more moisture leads to more absorbed energy and thus smaller reflectance (Jensen 2009). This naturally is problematic especially in field conditions with different moisture levels, as noted by Morgan et al. (2009). Spectroscopy with narrow bands, however, can find more precise wavelength

bands where these properties can be separated. Most soil properties, including SOC, can be studied in the VIS-NIR-SWIR (Visible - Near-Infrared - Short-wave-infrared) region from 0.4 to 2.5 µm (Ben-Dor et al. 1999).

Yang & Li (2013) suggest in their study that the main response of SOC is in VIS range but using the VIS-NIR spectrum produce better results in predictions. Ben-Dor (1999) agrees with this stating that VIS region is especially active in terms of SOC, but that SOC is also an important *chromophore* (material that absorbs incident radiation in discreet energy levels) in the whole NIR-SWIR region. Although the importance of the NIR and SWIR regions is widely approved when predicting SOC (Viscarra Rossel & Hicks 2015, Bartholomeus et al. 2008, Doetterl et al. 2013), moderate results have also been achieved with using more reduced spectral range of 450-950 nm (Jung et al. 2015). Viscarra Rossel et al. (2006) found good correlation between SOC and bands 410, 570 and 660 nm in the VIS spectrum, and no difference in using only VIS spectrum (400 - 795 nm) versus using only NIR spectrum (810 - 2400 nm) alone, although the combined use of VIS-NIR-MIR spectrum produced better results.

2.3.2.2 Multivariate methods for SOC prediction

Quantitative prediction of soil properties such as OC by its spectral characteristics is based on Beer's law, where it is assumed that there is a relationship between spectrometric response and the concentration of a material in the studied sample (Gobrecht et al. 2014). Even though Beer's law is the underlying theory behind spectral analysis, it is not regularly used as itself, rather the response of soil attributes are discerned from the spectrum with various statistical methods (Rossel et al. 2006). These include many multivariate methods from basic multiple linear regression (Nanni & Demattê 2006) to support vector machine regression (Peng et al. 2014), but the most used are partial least-squares regression (PLSR) and principal component analysis (PCA) (Vasques et al. 2008).

PLSR was found to be the most consistent and best in a comparison of different modeling methods for SOC by Vasques et al. (Vasques et al. 2008). The strength of PLSR and PCA comes in their ability to deal with large amount of highly collinear X variables (predictors)

by projecting them into new low-dimensional space of latent variables (Wold, S. & Sjöström & Eriksson 2001a).

3 Study area

3.1 Taita Hills

The soil samples were collected in Taita Hills in Taita Taveta County located in southeastern Kenya south from the equator by 3 degrees. The area can be roughly divided to dry lowland plains at an altitude from 400 to 1000 meters above sea level and highlands with diverse land cover from shrubland to indigenous montane forests up to 2200 meters (Pellikka, Petri KE et al. 2013).

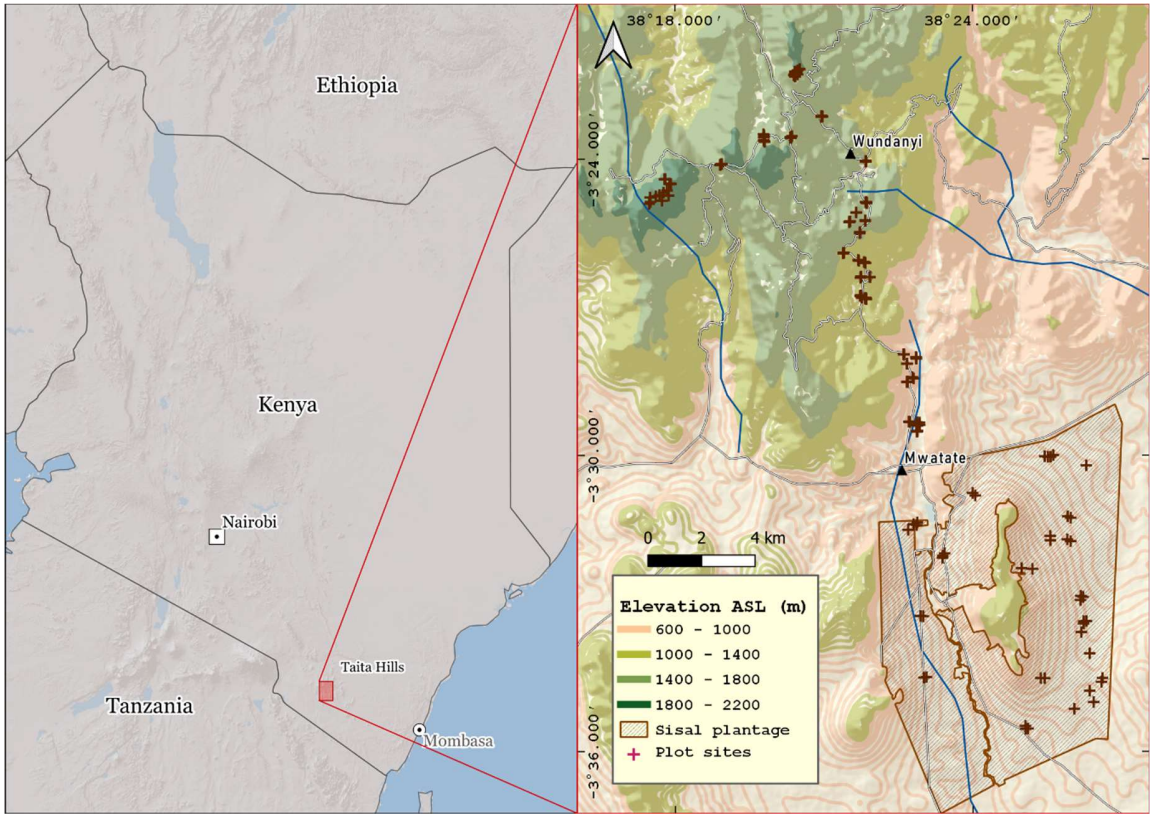

Figure 7. Research site in Taita Hills, Kenya and the study plots.

The Taita Hills area is surrounded by Tsavo National Parks in the lowlands. The hills are a part of the Eastern Arc mountain chain which dates to the Precambrian period and are classified as one of the world's 34 most important biodiversity hotspots with many endemic and endangered species of flora and fauna (Myers et al. 2000)

Following the influence of the Inter-Tropical Convergence Zone (ITCZ), the area experiences two rainy seasons, from March to June and from October to December. The annual rainfall ranges between 800-900 mm in the lowlands (Mwatate) to 1300-1400 mm in the hills (Mwalusepo et al. 2016). Due to favorable conditions, the hills are intensively cultivated, and remnants indigenous forests can only be found in small patches in the highland from e.g. Yale and Ngangao. The area of the indigenous forests has decreased by 50% from the year 1955 to 2004 although the total forest cover however has not decreased as drastically (only 2%) because of plantations of exotic species like eucalyptus (Pellikka, Petri K. E. et al. 2009).

The population of Taita-Taveta county has increased from 90 000 persons in 1962 to 340 000 in 2019, which is evidently the main reason for the loss of the indigenous montane forests (Pellikka, PKE et al. 2018), as the main livelihood in the area is agriculture. In the highlands, agriculture is intensive, but farm sizes are less than 0.5 ha, while in the lowlands farm size average is several hectares. (Autio et al. 2020). In the humid hills the soils are poor in nutrients, while in the dry lowlands the nutrient level is high (Jaetzold, Schmidt et al. 1983).

3.2 Soils in the area

Similarly as land cover and vegetation zones, soils in Taita Hills and the surrounding foothills and plains follow the patterns of topography. Following the soil classifications of the Farm Management Handbook of Kenya (Jaetzold et al. 1983), most parts of the lowland plains are covered with rhodic Ferrasols (PnF 1) (Figure 8), a well-drained, low fertility soil with dusky to dark red color. High fertility Fluvisols (AA 4) can be found in the lowlands around the alluvial plains. The foothill slopes form a zone of low fertility soils of type FU 2 around the hills. These are well drained dark red Ferrasols, Arenosols and Luvisols. The smaller hills in the lowlands are covered with a varying fertility HU 2 soil type, which varies between Regosols and Cambisols. The highlands of Taita Hills are mostly dominated by a moderate to high fertility Cambisols, Nitisols and Regosols of type MU 2. They are well drained and moderately deep reddish brown to brown soils with a distinct humic layer. The soils in the highlands vary between low fertility Acrisols, Cambisols and Ferrasols (UU 7) on the

moderately steep slopes to more fertile humic Rankers and Cambisols (UUC 3) in the montane forests.

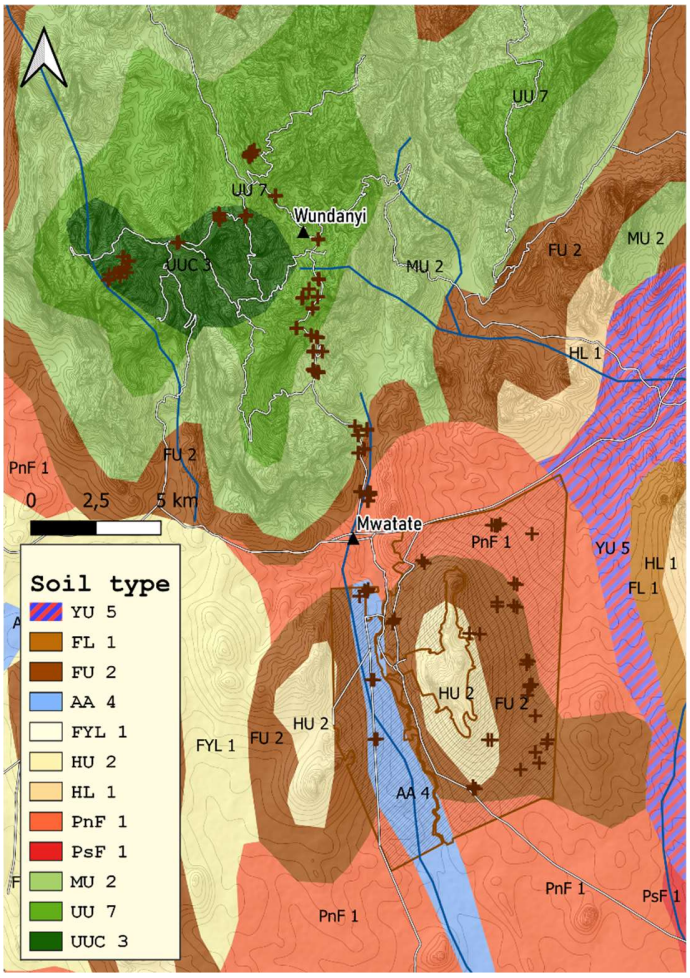

Figure 8. Soil map of the study area within the Taita Hills according to REF.

Topography and land cover are also the main factors affecting soil organic carbon (SOC) stocks in the area, as demonstrated in the research by Njeru et al. (2017). The general trend shows increasing amount of SOC from lower altitudes to higher, although there were varying values within plots in the same altitude level. Expectedly, largest SOC stocks resulted from the forests on top of the hills, where cold and wet soils limit the decomposition process allowing the accumulation of organic matter. These areas also have the largest above ground biomass (AGB) and above ground carbon density (AGC) in the region (Pellikka, PKE et al. 2018) leading to greater inputs of carbon to the soil. The plots

within mango trees (Mangifera indica) were found to have the least SOC according to Njeru (2017). In general, natural land cover, such as forests and bushlands, have greater SOC stocks although with more variance, compared to agricultural land use areas.

4 Materials & methods

4.1 Specim IQ

The hyperspectral data was imaged applying Specim IQ portable hyperspectral camera. The device weights 1.3 kilograms and with a size of 207 x 91 x 74 mm it is suitable for field measurements and dynamic imaging. The spectral range is 400-1000 nm with 7 nm spectral resolution and 204 spectral bands with almost 60 % overlap between adjacent bands. It has a line scanning pushbroom with CMOS sensor, while the image size is 512 x 512 pixels resulting to 512 x 512 x 204 data cubes where the first two dimensions are spatial and the third is spectral (Behmann et al. 2018).

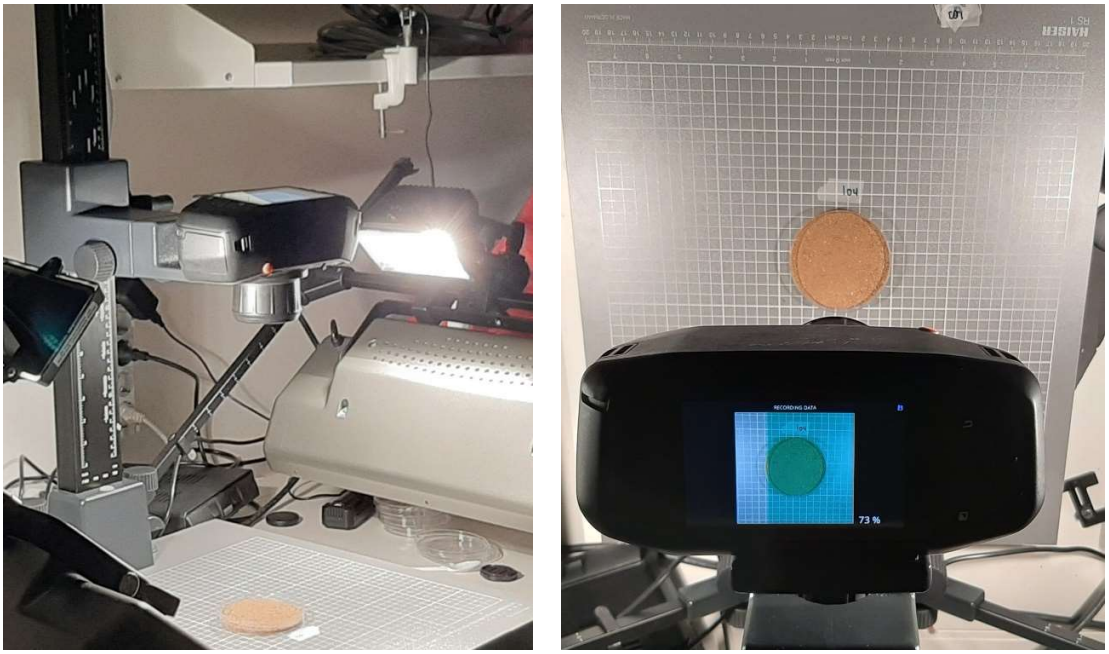

Figure 9 Specim IQ portable hyperspectral camera in action

IQ records three data cubes with one containing the raw intensity data and the two others are a dark frame and a white reference. Dark frame and the white reference are used for transforming the raw intensity data to reflectance to lose the effects of the illumination and other imaging conditions to get comparable measurements (Specim IQ manual.).

Previous research with IQ has focused on plant health and physiology monitoring (Barreto et al. 2020, Alt et al. 2020), plant phenotyping (Behmann et al. 2018) and geological applications (Derron & Jaboyedoff 2019, Kruglikov et al. 2019).

4.2 Field work

4.2.1 Sampling strategy

The sampling strategy has two principles. First, in order to study the levels of SOC in an altitudinal gradient and various land cover types, the samples were collected from lowlands around Mwatate at 800 m a.s.l. along a transect ending to montane forest of Vuria at 2100 m a.s.l. Second, for the development of the analysis method applying IS, an asset of samples of varying SOC levels were needed, which again required a comprehensive sampling of the soils from the area. The main application of the results is agricultural soil monitoring, thus samples from agricultural lands and soils were well represented in the sampling. SOC has been studied in the area previously by Njeru et al. (2017), and similar sampling strategy and plotting sites were used in this study enabling the comparison and validation of results of the carbon analyses. The plot sites (roughly one hectare areas) are located close to the main road from the town of Mwatate in the lowlands (~ 800 m) to the top of Vuria mountain (~ 2200 m) in an approximately 2km wide and 48 km long transect, with some additional plots at the Teita sisal plantation (Wachiye et al. 2020, Vuorinne 2020) close to Mwatate in the lowlands. 3 to 6 samples were collected from each plotting site to cover the variation of land cover in the site. The samples were classified to five land cover types from natural shrubland (n=17) and forest (n=32) to agricultural field (n=59), agroforestry (n=39) and sisal (n=44) with a total number of 191 samples (Figure 7. Research site in Taita Hills, Kenya and the study plots.Figure 7).

4.2.2 Sampling

Three soil cores from the depth of 20 cm were collected for each sample, within one square meter area using a soil auger. The three cores were sieved at the plot with a 2mm sieve net, mixed together and then stored in 0.5 l plastic bag (100-200 grams of soil per sample). The location of the sample was measured with Garmin GPSMAP 64x handheld navigator as the

approximate center of the three drill holes. Soil moisture and temperature were also measured, but these data were not used in this study due to low quality.

Figure 10. Research assistant Darius Kimuzi of Taita Research Station collecting and sieving the sample. Soil moisture was measured with Decagon Devices ProCheck reader.

The 10-day field campaign in late August 2019 was interrupted by rain for few days, hampering the soil moisture measurements as well as imaging of the samples in the field as moisture affects the soil color.

In order to transport the samples to University of Helsinki in Finland for carbon analysis and laboratory imaging each sample was reduced to 60-100 grams, because of the limits of the exporting license.

4.2.3 Field imaging

The samples were imaged in field conditions at the Taita research station within a few days from the field sampling under diffuse or direct solar illumination conditions. Specim IQ was

attached on a tripod at about 45cm in height over the sample placed on a 10cm diameter petri dish. All the samples from the sisal plantation were not imaged in field conditions due to time constraint.

IQ was used in the default recording mode with the white reference board placed with the sample in each image. Varying lighting conditions from partly cloudy sky to direct sunlight and occasional rains expectedly affected the field imaging. Sometimes the lighting changed during the 30-40 second scanning time from diffuse radiation to varying levels of cloud cover. Under direct sunlight the samples or the white reference board saturated easily, although good illumination enabled scanning times of few second quickening the imaging process. The above-mentioned conditions are normal situation in field imaging in Africa.

4.3 Laboratory work

4.3.1 Sample preparations

The samples were oven dried in opened plastic bags in 40 degrees Celsius at soil laboratory of the Department of Forest Sciences of the University of Helsinki. Most samples were in the oven for four days, but a few samples required six days to make sure that all the moisture was evaporated from all samples.

For the carbon and nitrogen analyses three small spoonfuls of soil from each sample was grinded using a stone mortar and muscle power and stored separately in small paper envelopes. These sub-samples were used for carbon and nitrogen analyses only, while the samples that were imaged were no further preprocessed after drying them in the oven. The grinding took 4 days since all the equipment needed to be cleaned with ethanol between each sample to avoid sample contamination.

4.3.2 Carbon and nitrogen analyses

Carbon and nitrogen analyses were conducted with a Leco CN828 automated dry oxidization analyzer. For each sample, about 300 milligrams of soil was weighted inside a tinfoil nugget and then placed in the analyzer. The analyzing time per sample was 2 minutes,

but up to 30 samples could be queued in the device so no time for preparations was needed between analyses. Analyzing all the samples took 2 days and the results were automatically exported to excel. Average SOC content of the samples in this dataset was 2.36%, with minimum of 0.29%, maximum of 14.6% and a standard deviation of 2.68 (Table 1). As expected, the spatial distribution of SOC values changed along the altitudinal gradient representing low SOC contents in the lowlands and high contents in the highlands, and especially in the forests (Figure 11).

Table 1. Carbon and nitrogen analysis results.

Parameter	**Mean**	**Min**	**Max**	**Stdv**
SOC (%)	2,364	0,292	14,600	2,268
N (%)	0,701	0,357	1,870	0,227

The measured SOC values per land cover type are shown in Table 2. Forest samples have the highest SOC values as well as variation in SOC content, whereas sisal samples have low average SOC and low variation, although the range in sisal samples is very similar with shrubland samples. Similarly with forests, field samples have also high SOC and high standard deviation.

Table 2. Measured SOC content in various land cover types.

		SOC (%)			
Plot type	**n samples**	**Avg**	**Min**	**Max**	**Stdv**
agroforestry	39	2,07	0,40	3,91	0,88
field	59	2,19	0,29	10,70	1,59
forest	32	5,51	0,51	14,60	3,43
shrubland	17	1,57	0,42	2,83	0,76
sisal	44	0,92	0,42	2,34	0,37
All	**191**	**2,37**	**0,29**	**14,60**	**2,27**

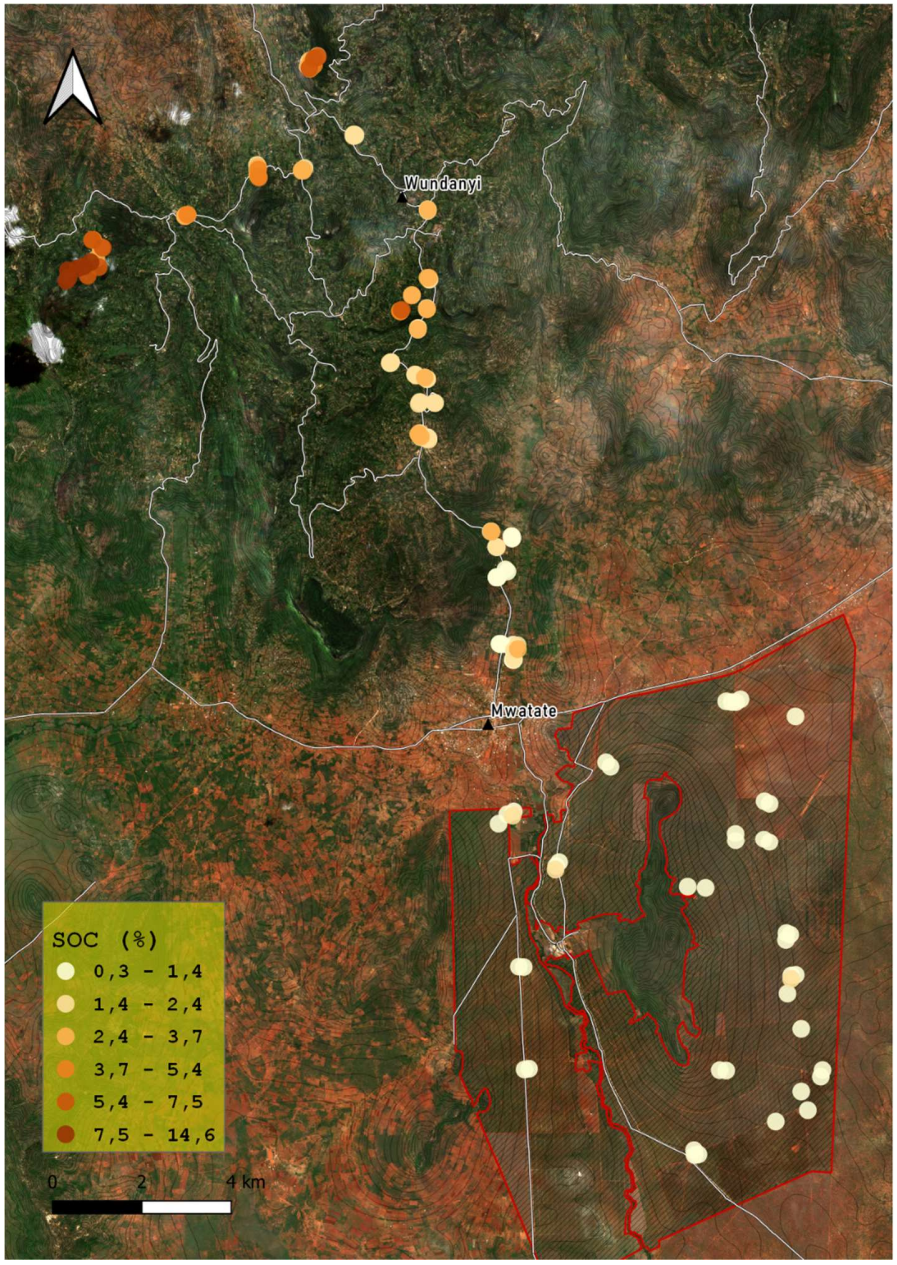

Figure 11. Soil organic carbon values resulted from the carbon content analysis. The carbon content is higher in the higher elevations, in which the vegetation cove is higher.

4.3.3 Laboratory imaging

The Specim IQ laboratory images of the soil samples were taken under supervised and stable conditions at the Department of Forest Sciences of the University of Helsinki. The setup included two 150 W halogen illumination sources and KAISER RS 1 imaging stand. The

camera was attached to the stand at an imaging height of 30 cm over the sample and the lights were set to 45 degrees angle on a 40-50 cm distance to the sample (Figure 12). The samples were imaged in a random order in 10 sample groups. The 10 samples were poured into 10 cm diameter petri dishes and imaged one at a time under the imaging stand. Between samples the petri dishes were cleaned to avoid sample contamination. Because of stable illumination conditions, Specim IQ was used in the custom white reference mode where the white reference for reflectance transformation is saved prior to the actual sample images instead of including it in every image and selecting the white reference area after each scan like done in the field imaging under varying illumination conditions. The saved white reference is used for all samples automatically thus removing one step from the imaging process. This method can only be used in stable lighting conditions in the laboratory.

Figure 12. The imaging set-up

4.3.4 Image pre-processing & extracting the spectra

Since the Specim IQ default recording mode saves reflectance transformed data in addition to the raw spectrum, no additional transformations were used. The data however required other pre-processing steps represented in Figure 13.

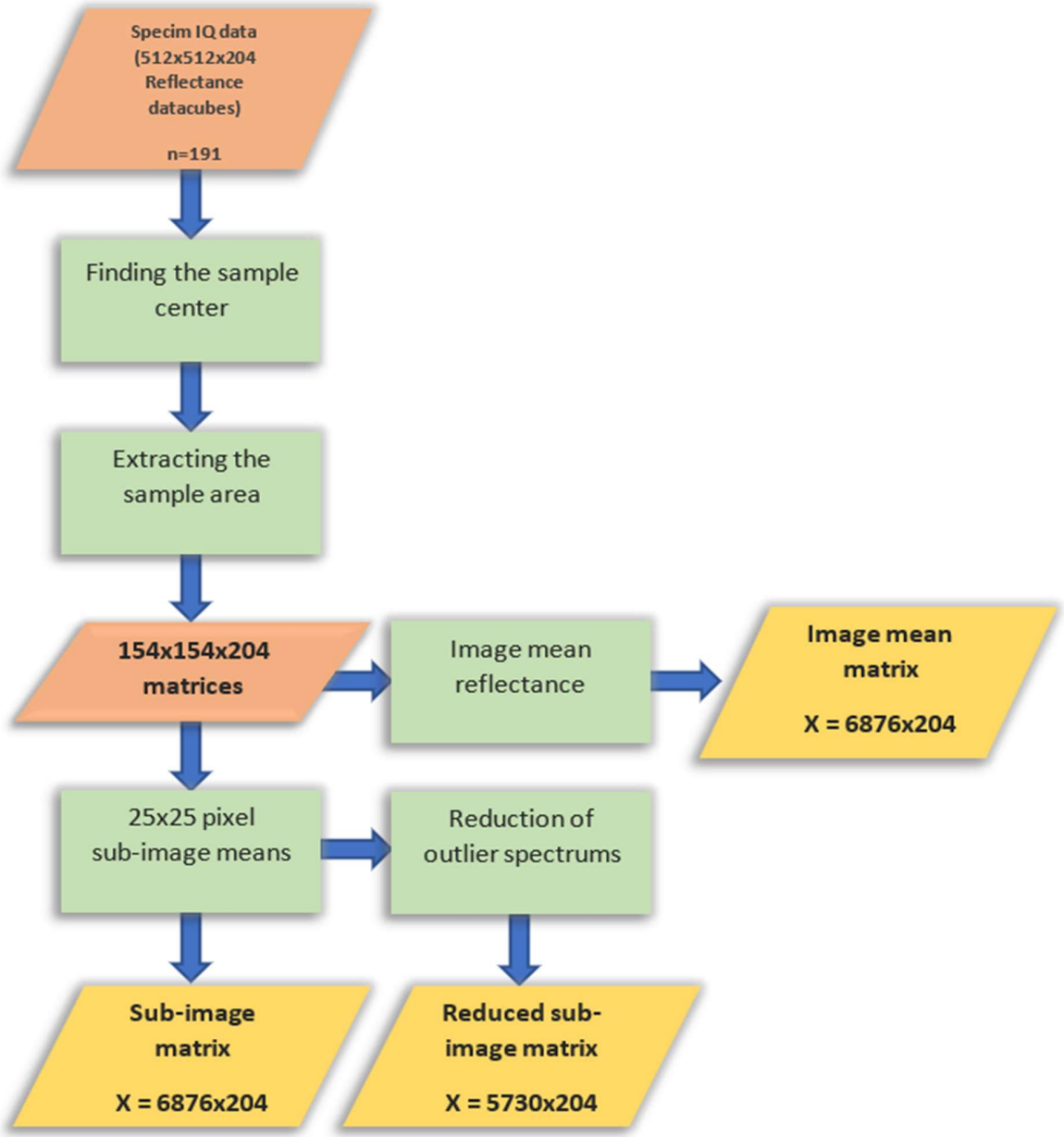

Figure 13. Image processing steps before analyses. Three datasets were formed, one with image mean reflectances, one with 25 x 25 pixel sub-image mean reflectances and one with reduced sub-image reflectances.

4.3.5 Cropping the sample area

The soil sample was cropped from the images to include only soil. However, as the samples were not placed in the same spot in each image, an image recognition HoughCircles algorithm from python library cv2 was applied to find the petri dish. This worked surprisingly well, but the images taken on 9.12.2019 gave more trouble in finding the sample circle. With the HoughCircles algorithm it was possible to extract the sample centers with enough precision to crop a 154x154 pixel rectangle inside the circular sample resulting in the end a 154 x 154 x 204 matrix for each sample (Figure 14).

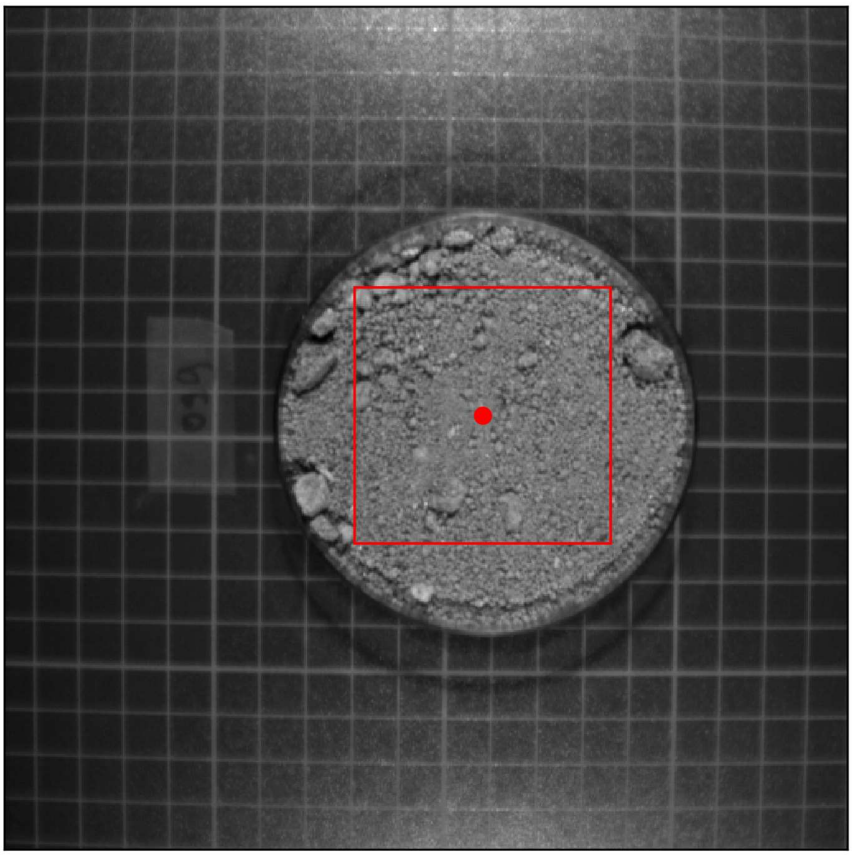

Figure 14. Visualization of original Specim IQ image of sample 099 and a 154 x 154 pixel rectangle representing the image area that is used for modelling

4.3.6 Sample average, sub-image and reduced sub-image datasets

With 154 x 154 spectrums gained with Specim IQ and only one corresponding carbon value from the dry oxidization analyses per sample, there is a large amount of redundant information in the spectrums since it is impossible to assess how a singular spectrum is

related to the SOC content. The spectral data was therefore averaged to three different datasets of mean spectrums.

The first data set contains the sample mean spectrums in a 191x204 X matrix where the spectral data is reduced to one spectrum per sample as a mean of the 154x154 spectrums. For this a corresponding y-matrix of carbon values was formed with a shape of 191x1.

Because a lot of information is lost in averaging the whole sample spectrum, two sub-image datasets were also generated.

For the second dataset the cropped sample images were segmented to 36 sub-images with a 6x6 grid where each grid cell covers 25x25 pixels (Figure 15). For this the images were reduced to 150x150 pixels for simpler division.

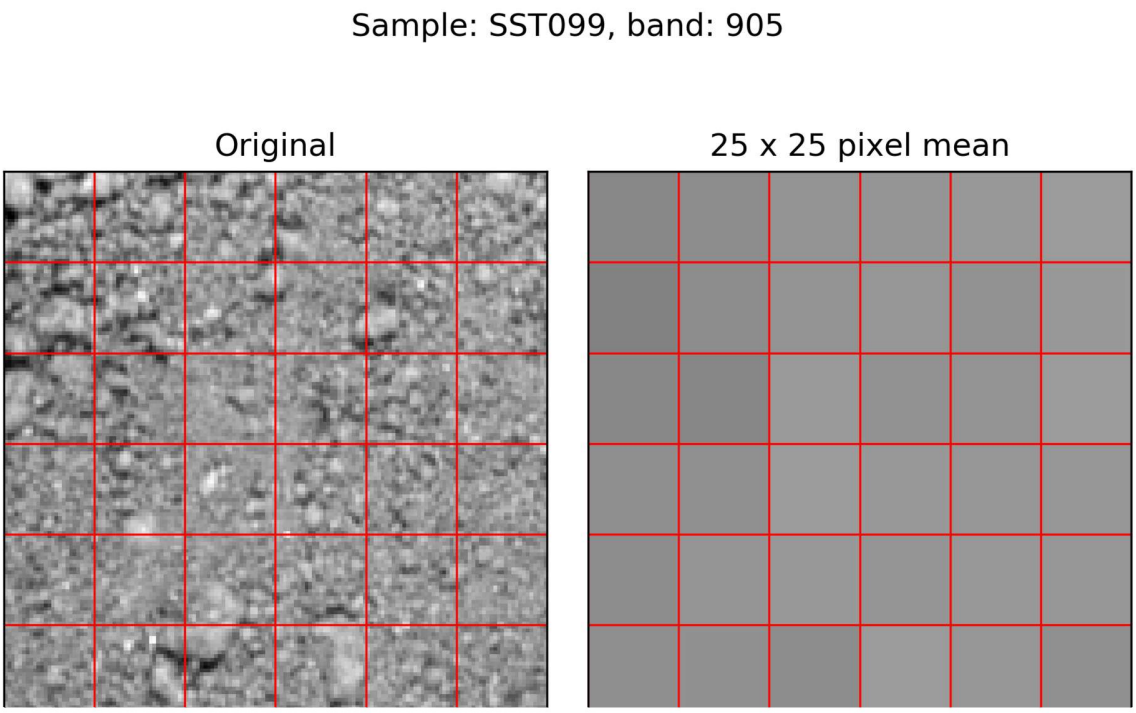

Figure 15. Original sample data of sample 099 and the averaged sub-image data visualized as band 905 nm

From the sub-images a 6876x204 X-matrix was formed where each sample is represented 36 times as a mean of the 25x25 pixel sub-image spectrum. For this a corresponding y-

matrix of carbon values with a shape of 6876x1 was also formed, where each carbon value is represented 36 times.

The third data set is a reduced version of the sub-image set, with 6 outlier mean spectrums dropped from each sample. The outliers were found by calculating the total intensity of each sub-image cell as the sum of the reflectance values *R* of all bands *n=204* (eq 1.). These intensities were sorted and three cells with the lowest and three cells with the highest intensity were dropped out from each sample. This resulted in 30 subsamples per field sample and as 5730x204 X-matrix (Figure 16). Again, a corresponding y-matrix of carbon values with a shape of 5730x1 was formed where each carbon value is represented 30 times.

$$I = \sum_{i=1}^{n} R_i \qquad \text{(eq. 1)}$$

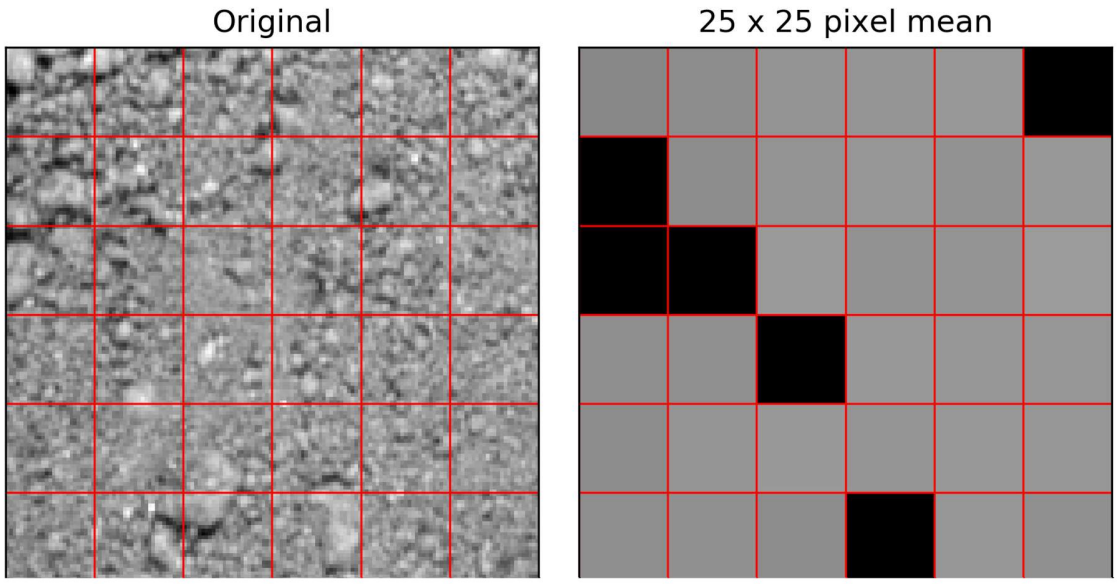

Figure 16. Original data and the reduced sub-image data visualized

4.4 Quality assessment

Quality assessment was performed only for the sample average dataset, since it's safe to assume that if the spectrum of an image would be disturbed, it would affect the whole image and not just part of it, as the imaging conditions were controlled and stable. Some assumably sensor-based disturbance was found from edges of the spectrum, and also in the blue bands of visual spectrum and the last infrared bands (Figure 17). No particular error was found from the data imaged on 9.12.2019, and the difficulties with this data during sample area detection was likely because of small focusing error while imaging. The disturbed edge bands were dropped out before running the models.

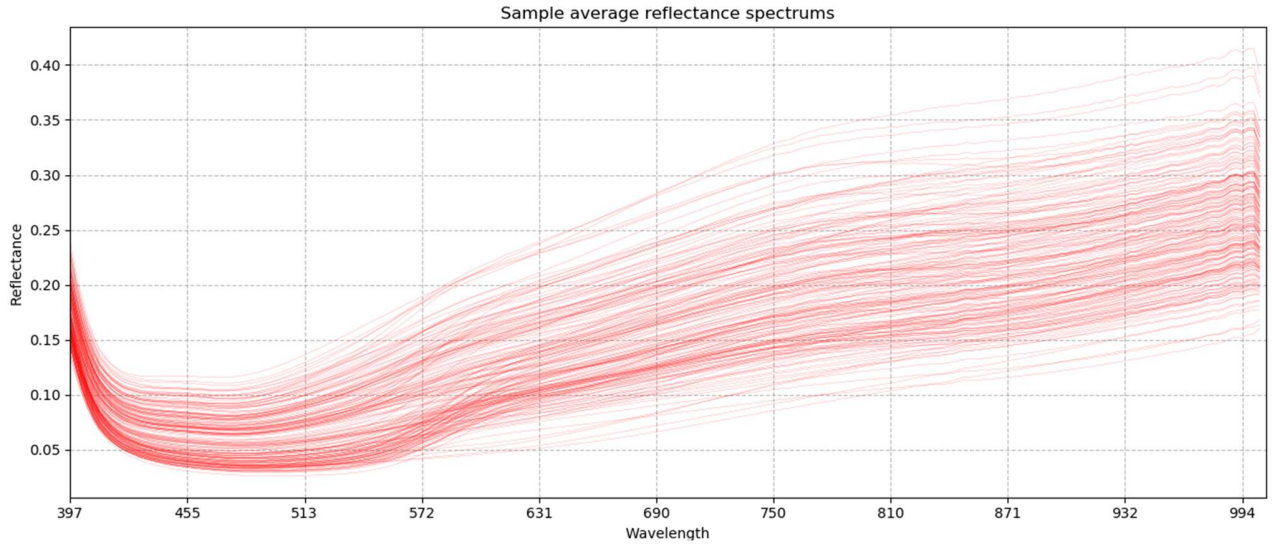

Figure 17 Mean reflectance spectra of all samples. The sudden rise in reflectance on the early blue bands as well as the last noicy NIR bands were dropped before running the models

4.5 Data pre-processing

The noisy bands from the edges of the spectrum were removed before implementing the models. After removing the first 12 bands in the blue region of visible spectrum and the last 10 bands of the NIR spectrum, the spectral range was reduced to 440-972 nm with 182 spectral bands.

The spectral data (**X**) was also centered with a *standard normal variate* correction (eq 2.) to get a mean of 0 and standard deviation of 1 for each sample spectra (Figure 18), by first subtracting the mean of each spectra from itself and then dividing it by its standard deviation:

$$X_i^{snv} = \frac{(X_i - \bar{X}_i)}{\sigma_i} \qquad \text{(eq 2.)}$$

As a result noise is reduced in the data and equal prior importance for each band in the analysis is given (Wold, S. & Sjöström & Eriksson 2001a).

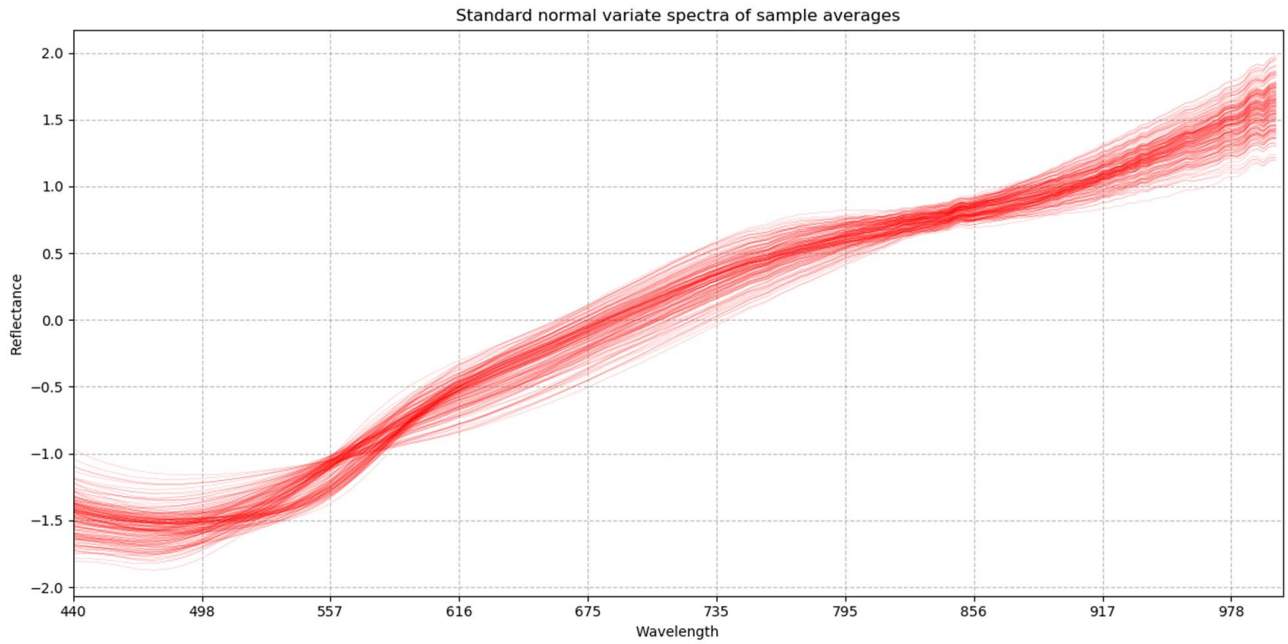

Figure 18 Standard normal variated spectra of sample averages with reduced bands.

Finally before running the models, the **X** and **Y** data was split to training and testing datasets by having 30% of the samples as test set and 70% as training set ending to 133 training samples and 58 testing samples. Since no stratified sampling method was used in splitting of data, the bearing of the split could be significant since the number of samples (n=191) is not very high. For this reason, 50 iterations of model building was conducted with different seeds for the splitting algorithm and the results are averaged from these.

4.6 Field data

Due to poor imaging conditions, the field imaging results had lot of disturbance especially in the NIR region (Figure 19). Some images were also completely unusable, and their spectra could not be extracted. In addition to missing images from some sisal plantation samples, the usable field data set included in the end 168 sample spectrums. Only one dataset was formed and tested from the field data, containing the sample average spectrums. The noisiest bands in the spectrum edges were dropped from the dataset leaving 165 bands to be used in the models.

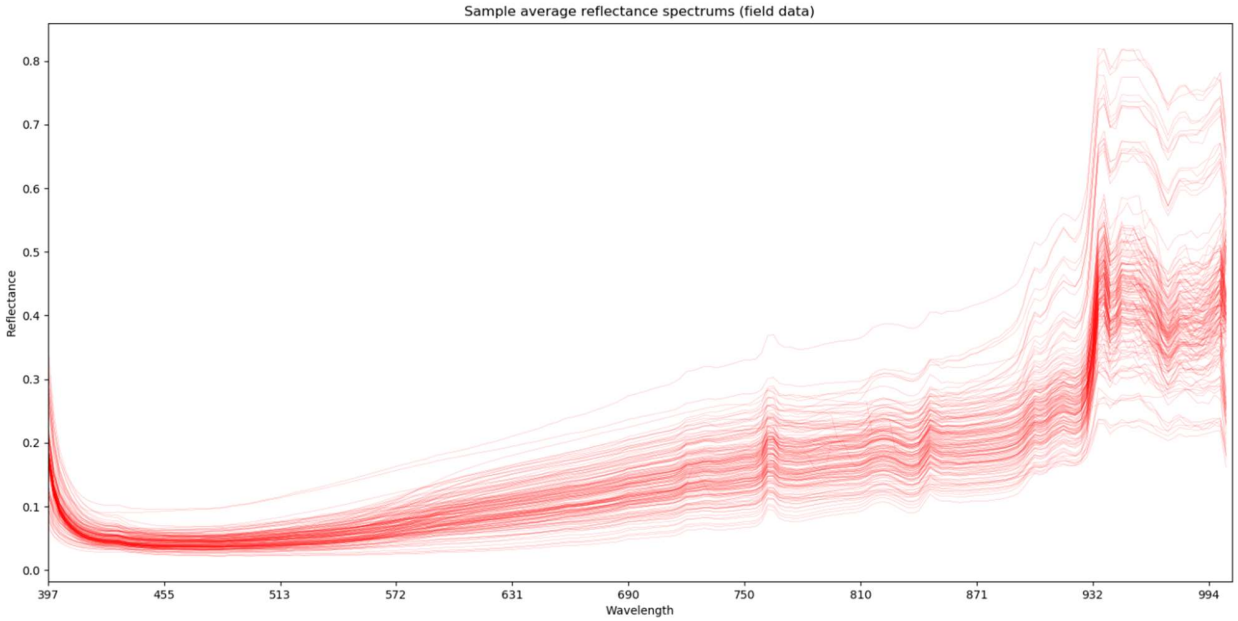

Figure 19. Average spectrums of the field samples.

4.7 Multivariate methods

Two different multivariate approaches were used. Partial least squares regression was chosen for its broad use and good performance in previous studies (Jung et al. 2015, Rossel et al. 2006, Vasques et al. 2008). Lasso regression was applied for its ability to handle multiple collinear features. No previous hyperspectral research concerning SOC or other soil properties utilizing lasso regression was found, although e.g Lazaridis et al. (2011)

concluded it to perform equally with PLSR in their tree mortality related research with satellite data.

All models used here start with matrices **X** and **Y** with the shapes of **X** = $(n \times k)$ and **Y** = $(n \times m)$, where *n* is the number of samples/observations, *k* is the number of explaining variables (bands in this case) and *m* is the number of response variables, which the model aims to predict. Since SOC is the only response variable here, **Y** is a column matrix of shape $(n \times 1)$. Both PLS and Lasso are linear models based on multiple linear regression (MLR) which in the end predicts **Y** as (Dyar et al. 2012):

$$y_i = \beta_0 + \sum_{n=1}^{k} \beta_n X_{in} \qquad \text{(eq 3.)}$$

Here β_0 is the intercept and β_n are the regression coefficients for parameters $\mathbf{X}_{in}$. The coefficients are determined by ordinary least squares (OLS) estimation.

All models were implemented in Python 3.7 using the models from scikit-learn library, see the model building workflow from Figure 20 .

4.7.1 Partial least squares regression (PLSR)

PLSR, introduced by Herman Wold in the 1970's (Wold, S. & Sjöström & Eriksson 2001b), is a standard tool in chemometrics and widely used in spectroscopy for its ability to handle multicollinear data with a large number of variables (Vasques et al. 2008).

PLSR is closely related to principal components regression (PCR), which is used to reduce the dimensionality of collinear data by finding new lower number of variables that represent the variability in the original data (Martens & Naes 1992). Whereas PCR forms the new components from features **X** so, that variance between each component is maximized, PLSR uses also the response data **Y** in the decomposition of **X** to find latent features that are important in predicting **Y**. So instead of using **X** with *k* variables to predict **Y**, PLSR finds a matrix **T** = $(n \times a)$, where $a < k$, and represents the number *latent variables* (LV), and uses **T** for the predictions. Each column of **T** ($\mathbf{t_1}$, $\mathbf{t_2}$,…,$\mathbf{t_a}$) is a linear combination of the original **X** variables. The **T** variables are formed so, that the covariance between the predictor variables **X** and response variables **Y** is maximized and the first variables explain

most of the variation. More on how the LVs are calculated and a detailed description of PLSR, see Martens and Naes (1992).

As previously mentioned, PLSR was implemented using scikit-learn Python libraries and functions. The optimal number of LVs was determined with an iterative leave-one-out cross validation process using the NIPALS algorithm (Wold, H. 1975) and the maximum number of LVs was set to be 40. For each number of LVs (1-40), a PLS model is created and then fed to sklearn cross_val_predict function along with the training **X** and **Y** data. The function runs a five-fold cross validation, meaning it splits the data into five equally sized (or close to equal) data sets and then uses four of these datasets to fit the PLS model, and predicts values for the remaining set. This is repeated until all samples are predicted once, and this is also repeated for all PLS models with 1-40 LVs. The optimal number of LVs is then determined by the minimum mean squared error (MSE) of the CV predictions. The used functions can be found in Appendix 1.

4.7.2 Lasso regression

Least absolute shrinkage and selection operator (Lasso) is a regularization technique for multiple linear regression that is used to balance the bias-variance trade off of the model by shrinking the coefficients and thus to reduce overfitting of the model (Tibshirani 1996). It is closely related to another regularization technique, Ridge regression, with the difference that Lasso can shrink coefficients to be exactly 0 and consequently performs also variable selection.

Lasso shrinks the coefficients by setting a penalizing term to the residual sum of squares function (RSI) that is the basic estimation function for ordinary least squares regression (OLS). With standardized predictors x_{ij} and response values y_i when i = 1, 2, …, N and j = 1, 2, …, k, lasso aims to find coefficients β_j to minimize

$$\sum_{i=1}^{N}(y_i - \sum_j x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{k}|\beta_j| \qquad \text{(eq 4.)}$$

, where $\alpha \sum_{j=1}^{k}|\beta_j|$ is the penalizing term that is used for shrinking the coefficients. This penalizing term is controlled with the α (alpha) and when it's set to 0, the algorithm is

equivalent to RSI. Conversely, if alpha is increased infinitely, all coefficients will shrink to 0. For in-depth description of lasso, see Tibshirani (1996).

Lasso regression was implemented with Python and scikit-learn, as previously mentioned. The optimization only requires finding the optimal alpha value, which is determined similarly to the PLS implementation, that is by iteratively testing different alpha values with the cross_val_predict function and choosing the best value for alpha by the minimum MSE.

4.8 Sub-image prediction considerations

As the sub-image predictions end up with 36 or 30 estimates for carbon per sample, these predictions are averaged as a mean of these estimates to gain one prediction for each sample, since there is also only one reference value for each sample. For this it is also necessary to handle the train-test split in a way, that all sub images of one sample are placed in the same dataset, so that no sample is split between training and testing sets and the predictions are based on the whole sample and not just part of it.

4.9 Model evaluation

The model evaluations are based on root mean squared error (RMSE) and R^2 score. These are widely used and make it easy to compare results with other research (Rossel et al. 2006).

RMSE (Eq 4.) indicates the error between the predicted values and the measured values, it is always positive with values closer to 0 meaning a better performing model, and a value of 0 would mean a perfect model. It has the same unit as the values under evaluation and is thus easy to interpret.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 / \, n} \qquad \text{(eq 5.)}$$

R^2 score or coefficient of determination describes the proportion of the variance that is explained by the model in the dependent variables. It varies between 0 and 1 with higher values indicating better model. Mathematically it is the ratio of explained variation in **y** and the total variation in **y** (eq 5.)

$$\sum_{i=i}^{n}(\hat{y}_i - \bar{y}_i)^2 \;/\; \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \qquad \text{(eq 6.)}$$

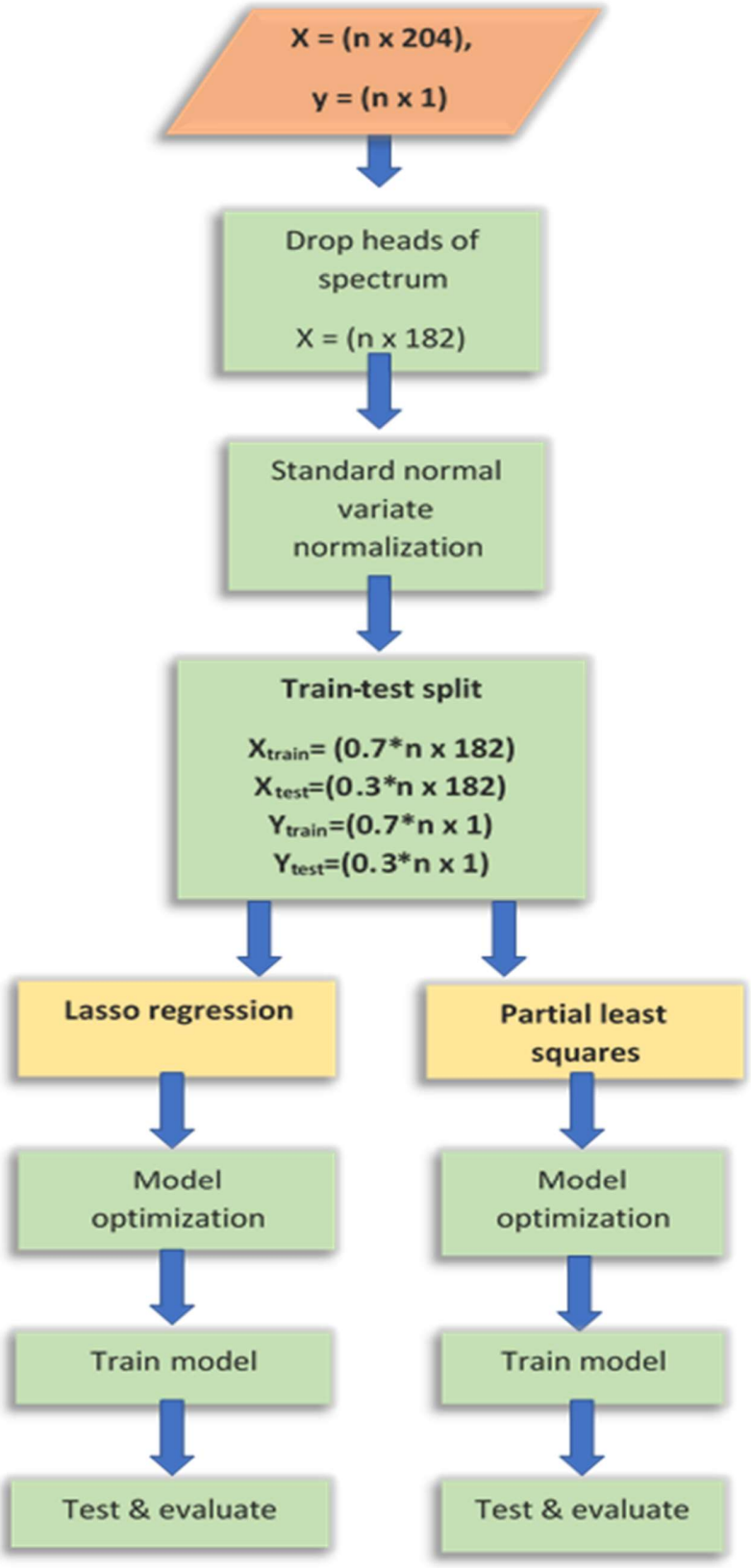

Figure 20. Model building workflow

5 Results

5.1 Sample average spectrum models

5.1.1 Overall model performance

In general, the sample average models were able to predict SOC fairly well and with very similar results, although PLS results were slightly more constant (Table 3). The average root mean squared error (RMSE) of measured and predicted SOC for PLS models was 0.978 and 0.967 for lasso with standard deviations (SD) of 0.124 and 0.136, respectively. Average R^2 scores were 0.77 for both PLS and lasso, with standard deviations of 0,06 and 0,11 respectively. Maximum R^2 scores were 0.89 for both models while the minimums varied a little more with 0.56 for PLS and 0.41 for lasso.

Table 3 Performance statistics of SOC predictions and optimization results of average spectrum models after n=50 iterations with different train-test sets. Ncomp is the number of latent variables found by PLS regression, nbands is the number of significant bands found by lasso regression. Ncomp and nbands averages are presented as mode.

	PLS regression			**Lasso regression**			
Statistic	**RMSE**	**R2 score**	**ncomp**	**RMSE**	**R2 score**	**alpha**	**nbands**
avg/mode	0,978	0,77	15	0,966	0,77	0,0000562	29
stdv	0,124	0,09	2	0,136	0,11	0,0006025	8
min	0,682	0,56	10	0,754	0,41	0,0000316	8
max	1,331	0,89	16	1,310	0,89	0,0017783	42

5.1.2 Model optimization results

The average number of components after the optimization for PLS models was 15, with relatively little variation (SD 2) and a minimum of 10 components (Table 3).

The alpha value in lasso regression was almost constantly optimized to 0,0000562 ending up with an average number of 29 selected significant bands (with a regression coefficient higher than 0). With lasso bands there was however somewhat higher variation compared

to PLS components, with SD of 12 and a maximum number of 42 bands and a minimum of 8 (Table 3).

5.1.3 Indicative wavelengths

As can be seen from Figure 21, lasso regression band selections were very constant especially with the high absolute value coefficients. Most important bands (ones with the highest absolute coefficient values) were found from the green/red transition region of the VIS spectrum around 600, 650 and 670 nm. In the infrared region there are a few constant important bandwidths around 750-775 nm and around 930 nm, but also quite a lot of inconsistency in the 930-970 nm region. In addition, there are few bandwidths of minor importance in the green and blue region of VIS spectrum and in the NIR at around 890 nm.

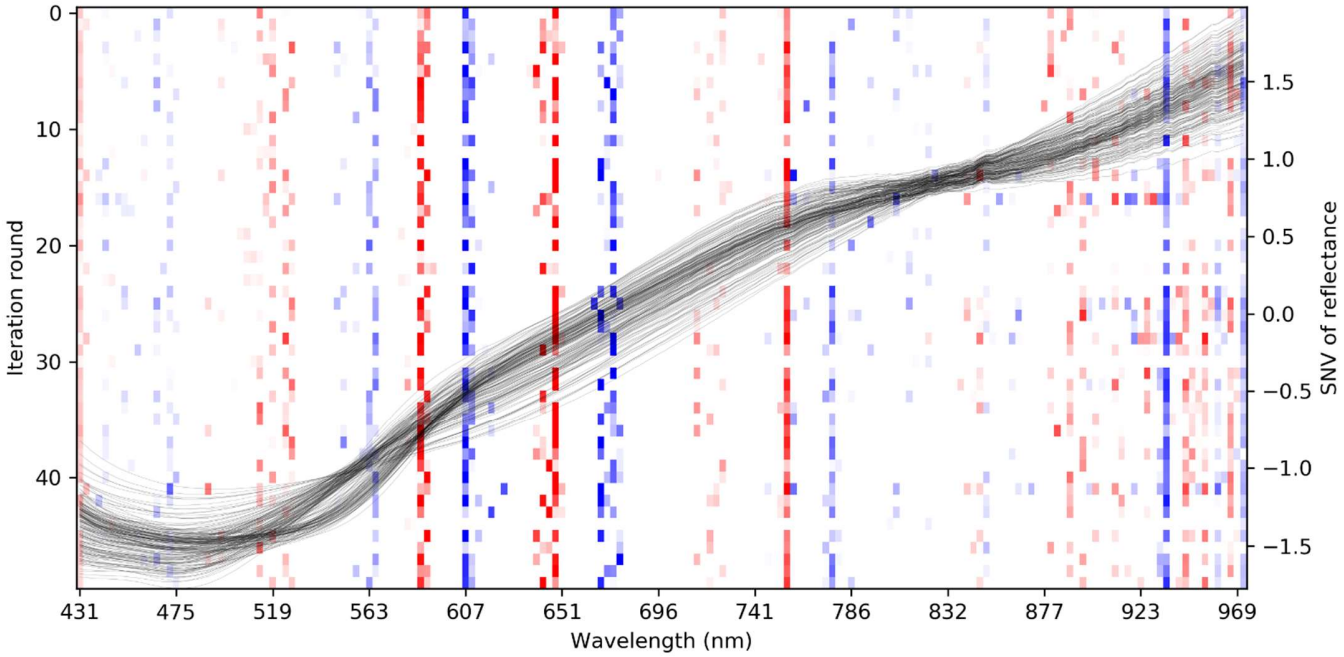

Figure 21. Lasso band selection results for all 50 iterations of train-test sets. Bright red indicates low negative coefficient values whereas bright blue indicates high positive coefficients, shallow colors indicate coefficients with low absolute value and white areas are bands with 0 coefficient. Coefficient value represents the correlation magnitude of the band with SOC in these models.

Coefficients extracted from the PLS models show similar results although the plot is not as interpretable as with lasso (Figure 22). These coefficients do not represent the actual components and are not used in prediction, but they tell about the relation between the **X** variables and the **Y** found by the model. While the plot is much noisier than Figure 21 of the lasso bands, same patterns and important wavelengths do stand out around the red region and in the early NIR bands.

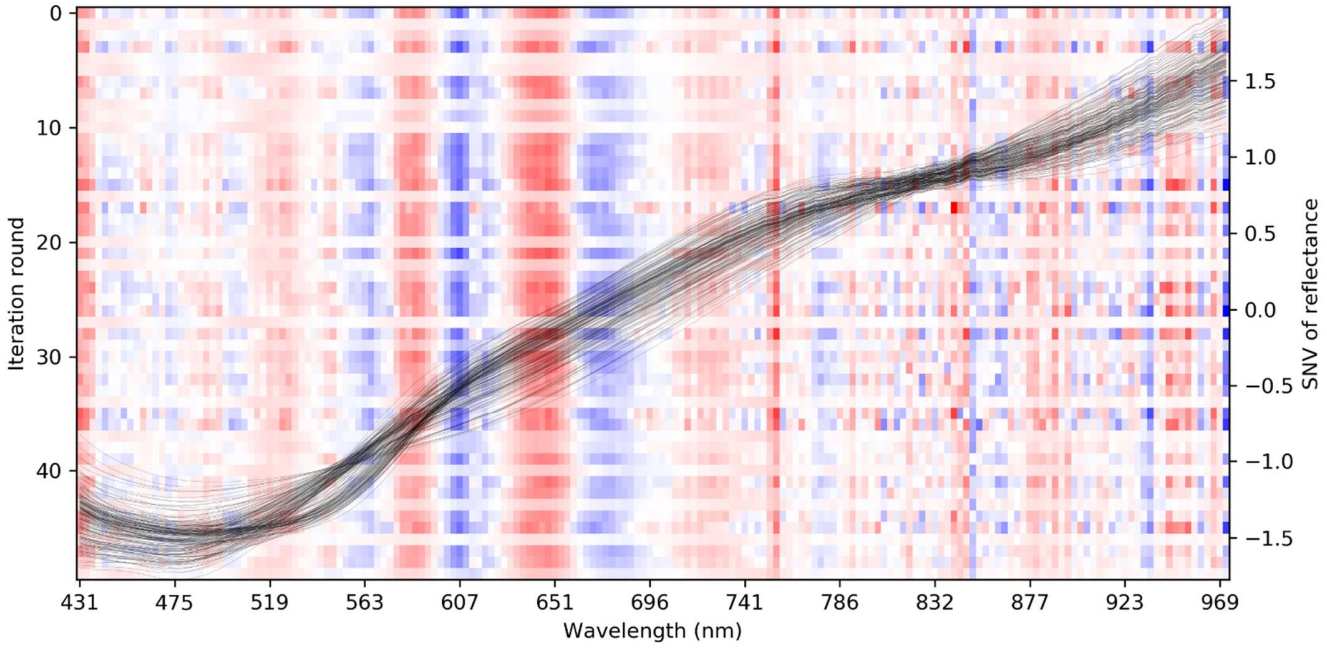

*Figure 22. Coefficients extracted from the PLS models (n=50) indicating the relations between the **X** (spectral bands) and **Y** (measured SOC) data found by the models. Bright red indicates low negative coefficient values whereas bright blue indicates high positive coefficients and shallow colors indicate coefficients with low absolute value. Coefficient values represent the magnitude of correlation with the band and SOC found by the model.*

5.2 Sub-image models

Optimizing the sub-image (SI) and reduced sub-image (RSI) models required very long processing times with the equipment in hand, so these models were only tested with three train-test splits. These splits were chosen by sample average model iteration results so that one poorly performing, one averagely performing and one well performing split was used.

In Figure 23 are the prediction results of the average performing split with PLS and Lasso, and all three datasets (sample average, SI and RSI data).

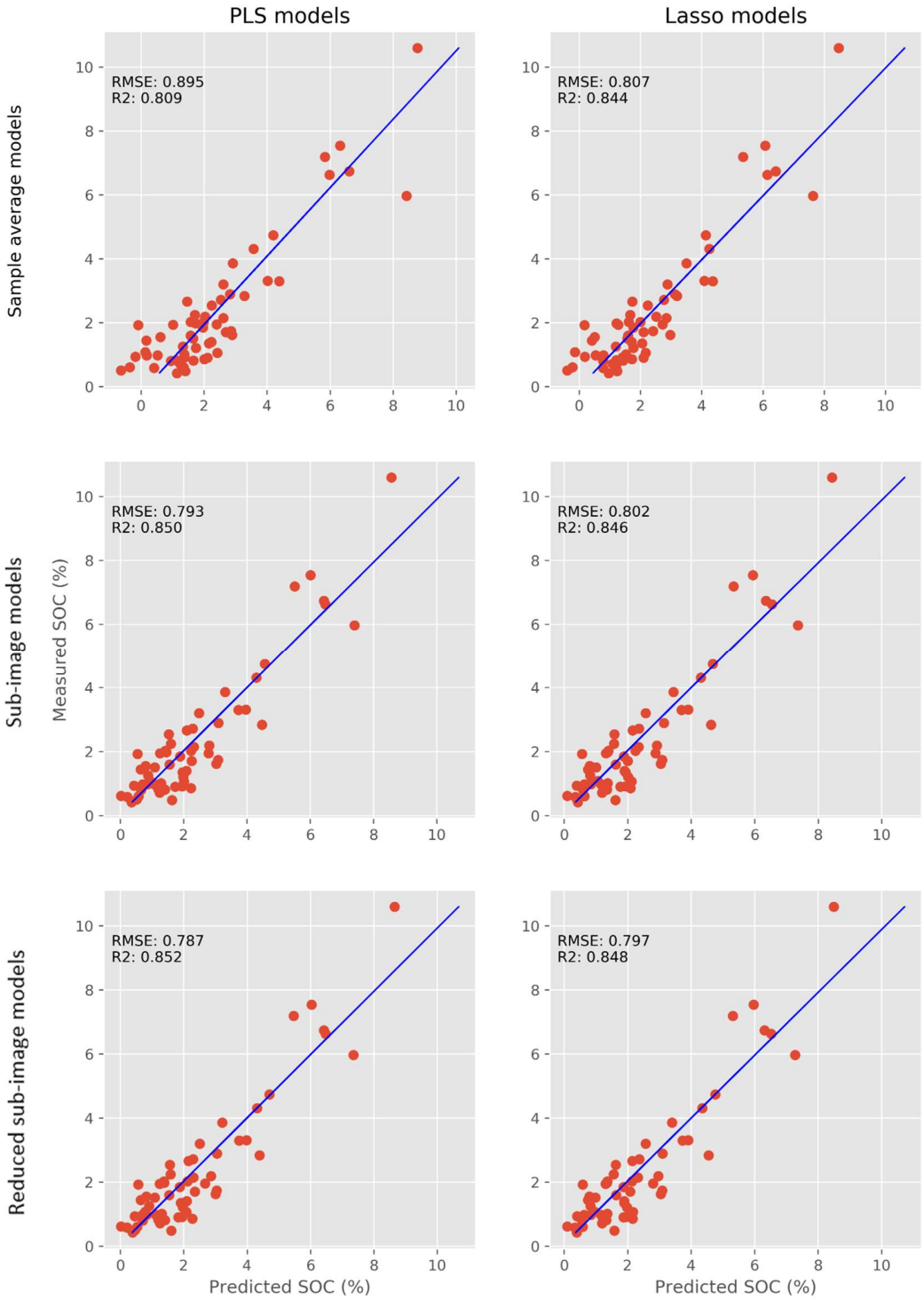

Figure 23. Model performance comparison of PLS and Lasso with sample average, subsample and reduced subsample datasets for predictiong SOC. Number of components used by PLS models here were 14,18 and 18 for sample average, sub-image and reduced sub-image model respectively. In the same order, number of selected bands for lasso models were 28, 81 and 76.

As can be seen from the plots and Table 4, using the sub-image data and reduced sub-image data in model optimization had a positive effect on the results in this case, especially with PSL models, raising the R^2 score to 0.85 and a 0.1 improvement in RMSE. Lasso models performed very similarly with all datasets, although with very slight improvements when using the subsample data.

Table 4. SOC prediction statistics of sample average, sub-image (SI) and reduced sub-image (RSI) models. First row is the laboratory measured SOC (%) statistics for the n = 58 test set

	SOC (%) prediction statistics				Model evaluation	
Model	**Mean**	**Min**	**Max**	**Stdv**	**RMSE**	**R2 score**
Measured SOC (%)	2,305	0,422	10,600	2,045	1,000	0,000
PLS	2,262	-0,631	8,772	1,975	0,895	0,809
PLS-SI	2,325	0,022	8,557	1,860	0,793	0,850
PLS-RSI	2,326	0,027	8,652	1,865	0,787	0,852
Lasso	2,268	-0,392	8,469	1,881	0,807	0,844
Lasso-SI	2,329	0,092	8,436	1,848	0,802	0,846
Lasso-RSI	2,326	0,094	8,495	1,849	0,797	0,848

All models seem to underestimate high SOC values while in the lower values the error is more evenly distributed as slight under- and overestimations. Sample average models tend to underestimate the predictions a bit more compared to SI and RSI models, although the STDV is closer to the measured data. SI and RSI models are free of the issue of predicting negative values and the prediction means for these models are really close to measured data mean (table 4.). Altogether there are no distinct outliers and all predictions are in a clear linear order.

Prediction accuracy varied somewhat between samples from different plot types (agroforestry, field, forest, shrubland, sisal) with forest type samples being the most difficult to predict (Table 5). This is likely due to forest plots also having the largest variation and largest individual values of SOC. On average, agroforestry and shrubland samples are overestimated whereas samples from other plot types are underestimated. Shrubland

samples have the best predictions results, it's also however the smallest class with only 5 predicted samples.

Table 5. Measured and predicted SOC values (reduced sub-image PLSR) per plot type. Forest type plots have most difficulties in predictions with highest average absolute error, this can be explained with also the highest standard deviation of measured SOC values

		Measured SOC(%)		Predicted SOC(%)		
Plot type	**n samples**	**Avg**	**Stdv**	**Avg p**	**Stdv p**	**Avg abs error**
agroforestry	39	2,07	0,88	2,13	1,16	0,61
field	59	2,19	1,59	2,12	1,48	0,67
forest	32	5,51	3,43	5,49	2,82	1,16
shrubland	17	1,57	0,76	1,89	1,03	0,52
sisal	44	0,92	0,37	0,89	0,65	0,56
All	**191**	**2,37**	**2,27**	**2,38**	**2,15**	**0,70**

With the poorly performing train-test split and well performing train-test split there was no distinct impact in predictions using the subsample datasets, seeming that these results are affected highly by division of the dataset.

In all cases with the subsample datasets the number of significant bands with a regression coefficient over 0 found by lasso models was much higher compared to sample average models. Models presented in Figure 23 and Table 4 reduced the number of bands to 81 and 76 for SI model and RSI model respectively, compared to 28 bands used by the sample average model. The number of components in the PLS models were also slightly higher with 18 components in the SI/RSI models compared to 15 components in the sample average model. This same pattern of higher number of components especially in lasso regression was also found in the models tested with different train-test splits.

5.3 Field data models

The field data models did not produce any reliable results as was expected due to quality of the data and thus no great effort was taken to develop them any further. The average R^2

scores were substantially below 0.5 for PLS and lasso models for sample average data after 50 iterations of model optimization with varying train-test splits. Average RMSE was 1.71 and 1.65 for PLS and lasso respectively also indicating the low performance of these models (Table 6).

Table 6. Performance statistics of SOC predictions and optimization results of field average spectrum models after n=50 iterations with different train-test sets. Ncomp is the number of latent variables found by PLS regression, nbands is the number of significant bands found by lasso regression. Ncomp and nbands averages are presented as mode.

	PLS regression			**Lasso regression**		
Statistic	**RMSE**	**R2 score**	**ncomp**	**RMSE**	**R2 score**	**nbands**
avg	1,710	0,35	7	1,657	0,41	17
stdv	0,294	0,31	2	0,436	0,28	6
min	1,193	-0,95	2	1,131	-0,64	3
max	2,490	0,66	12	3,774	0,66	32

The measured field moisture readings were not utilized in this work because some of the measurements were not very reliable due to rains during field work. The moisture content was also likely reduced before the imaging since there was usually at least a few days gap between the sample collection and the imaging.

6 Discussion

6.1 Modelling performance

The overall performance of the models was good compared to other studies using small scale imaging spectrometers for soil carbon estimation. Jung et al. (2015) tested the feasibility of a snapshot hyperspectral camera in SOC estimation in similar conditions and with similar methods using various PLS regression models and also with datasets of sample average spectrum and sub-image spectrums. R^2 scores in Jungs research varied from 0.51 with sample average models to 0.69 with sub-image models, which are significantly lower compared to results in this study. In a different kind of approach of imaging a soil profile in laboratory conditions, Steffens et al. (2013) obtained similar results with a R^2 score of 0.81 for soil carbon with PLSR, while the RMSE value was considerably higher at 2.62. Results of this study are also comparable with results gained in various studies in the last two decades, listed by Viscarra Rossel et al. (2006), even though most of these studies have utilized the whole VIS-NIR (until 2500 nm) with spectrometers.

6.1.1 Potentiality of lasso regression

PLS regression has already proven to be a feasible modeling technique for SOC prediction by a large number of studies (Vasques et al. 2008), but the results of this study demonstrate the potentiality of lasso regression as an alternative model. There was no significant difference in the accuracy of the models although lasso did not show any improvement in performance when the sub-image models were used. Similar results have been gained by Dyar et al. (2012) studying elemental composition determination of geological samples using spectral data and lasso and PLS, with lasso moderately outperforming PLS.

The benefits of using lasso are its relatively easy implementation and overall simplicity in comparison to PLS. The results are also more interpretable because the regression coefficients are in direct relation with the original variables enabling intuitive examination of important wavelengths or what ever the variables might be.

There are also some downsides with lasso, e.g. the optimization of the models was very exhaustive when using alphas smaller than 0.0001 (and the optimal alpha was almost

always below this), which limited the possibility of testing the models. Running the 50 iterations of sample average models took about 10-12 hours where most of the time was contributed to lasso optimization since PLS models were constantly optimized in less than one minute. The optimization time was further multiplied when introduced to larger sub-image datasets which brings us to the second downside of lasso, which is its challenge to find the significant variables when using larger datasets or more noisy data. This could be seen in the results as a multifold number of selected variables and low impact on the results. Dyar et al. (2012) also mentions the possibility of arising differences between PLS and lasso when using larger datasets and the results gained here seem to indicate that PLS could be more capable of finding the important information from large noisy datasets than lasso.

6.1.2 Sample spectrums and significant wavelengths

Most important wavelengths for SOC prediction were found around the red region of VIS spectrum by the models used in this study (Figure 21 and Figure 22). Additionally, few bands in the NIR region were constantly selected with high coefficients and some minor bands with lower coefficient values were found in the green region, although not with the same confident. Few studies have used the same spectral range and shared detailed results on the selected wavelengths, but similar findings have been reported e.g by Bartholomeus et al. (2008) about the 600 nm region, noting the 640-690 as the most important one. Viscarra Rossel et al. (2006) also mentions wavelengths in the same region to be significant for SOC prediction. No real connections between the important NIR bands discovered in this research was found with other studies, this could be because most studies have used the whole NIR region until 2500 nm and more important bands in the longer wavelengths are overshadowing the ones found here. The bands in the NIR edge of the used spectrum also showed a lot of inconsistency, which indicates that these bands may still be disturbed and should be left out of the models altogether.

Better understanding of the important wavelengths found in this work would require more research beyond the scope of this study, but the distinct similarity of the PLS coefficient plot (Figure 22) and the lasso coefficient plot (Figure 21Figure 22) signal that at least for the data used here, these wavelengths are important for SOC prediction.

6.1.3 Modeling and data considerations

Although the modeling results in this work have been promising, some considerations should be noted when assessing their robustness. To begin with, while the R^2 scores have been very comparable, the RMSE scores are slightly higher as compared to results of former studies listed by Viscarra Rossel et al. (2006). Average RMSE of the sample average models was close to 1 (Table 3) with some improvement when using the sub-image models (Figure 23), nonetheless meaning that precise small-scale predictions are not very reliable, which can be problematic when assessing e.g the critical limit of SOC concentration of 1.1 % (Lal 2004) or temporal short-term changes in SOC concentration of a field. In the following sections some factors that may have had an impact on the modeling results are discussed.

6.1.3.1 Sampling strategy

The samples were collected along a transect with an altitude variation from 800 m to 2200 m. The plots were distributed in various land cover and land use areas from agricultural fields to shrublands and montane rainforest. Due to the sampling strategy the samples had high variation in features from texture and color to carbon content. This is possibly a factor contributing to the good modeling results, as the various sample types are easily separated whereas small scale differences in SOC content of samples spatially close to each other are hard to recognize.

The samples were also not evenly distributed along different altitudes and land cover classes and e.g. the number of samples from the sisal plantation in the lowlands was quite large. This may have had some effect on the models as these samples were quite similar.

These models were only tested on this regional dataset and there is no guarantee of their applicability outside of this dataset or region.

6.1.3.2 Distribution of the measured C values

The variance of C values was also relatively high with a minimum of 0.2 % and a maximum of 14.6 %. This variance is highly correlated with the increasing altitude and change of the land cover from lowland savannah to tropical forests in the highlands, thus again leading to easy separation of high SOC value samples from low SOC value samples, but difficulties in

recognizing small-scale differences. This high sample variation is most likely the main reason for the high R^2 scores and the relatively weaker RMSE scores.

6.1.3.3 The number of samples

As could be seen from Table 2, the modeling results can vary greatly depending on how the data was split for training and testing sets. Although the average scores and the standard deviation were not bad, the minimum and maximum scores are very likely to be results of testing and training sets that fail to represent the whole dataset because of insufficient number of samples.

This problem could be partly tackled with stratified sampling methods where it would be certain that different kind of samples are represented in both datasets. However, the better way to resolve this issue would be a higher number of samples to further improve and stabilize the results.

6.1.3.4 Sub-image and reduced sub-image data

Using the sub-image and reduced sub-image models seem to have positive effect on the models, however these models were not tested as thoroughly as with the sample average data.

The downside of these datasets is the increasing complexity of the models as the amount and variance of different spectrums multiplies but the distribution of response variables (measured SOC values) stays the same. This leads to several diverse spectrums pointing to the same response value and may confuse the models. Especially the lasso models seem to be affected by this as seen as a growing number of significant bands.

On the other hand, these datasets manage to better capture the actual spectral variance of the samples and reduce the effect of outlier spectrums caused by e.g. shadows or plant debris that has remained in the samples. This was well seen as the relatively high performance improvement of PLS model.

Jung et al. (2015) used very similar approach with also improved results only they used the sub-image datasets just for predictions with models calibrated on sample average data, and

calculated the SOC as a mean of the second and third quartiles of the predicted values. Some test runs were carried out with sample average calibrated models during this work also, but better performance was achieved with re-optimization of models with the SI and RSI data and the calibration results mostly differed from the ones gained with sample average data.

6.1.3.5 Other possible model improvements

The models could be further adjusted with several ways from data improvements to calibration fine-tuning, as could be done e.g. for the alpha values of the lasso models, that were selected quite simply by trying out a limited set of values and selecting the best performing one.

Some studies (Jung et al. 2015, Peng et al. 2014) have also used various wavelength selection methods before running the models to reduce the data dimensionality. This was also considered in this study, but was left out since both PLS and lasso are used for dimension reduction/variable selection themselves and the effects of these methods in similar research setup in Jung et al. (2015) were not significant.

6.2 Applicability of the methods

As shown in this study, Specim IQ is at least technically suitable for estimating SOC at a moderately good level in a controlled environment. In the following section its usability for this purpose in practice is more closely assessed.

6.2.1 Laboratory application and improvement suggestions

Greatest advantage of using Specim IQ is the low sample preparation requirement in comparison to automated C/N analyzer. Only preparations made in this research were sieving the samples through 2mm sieve during sample collection, and oven drying the samples as pretreatment before the carbon analyses. While using the automated C/N analyzer is very easy and fast, the required grinding was slow, laborious physical work. According to Morgan et al. (2009) grinding the samples doesn't affect the results with reflectance spectroscopy and the good results gained here refers the same.

Still, as they are, the methods used in this study do not fully utilize the spatial aspect of the data and offer no advantages over using point spectrometer which usually have better spectral range and are better tested. To more utilize the imaging aspect of IQ, several samples could be imaged simultaneously, as was done by O´Rourke et al. (2011). This would speed up the sample analyzations with a price of less data points (pixels) per sample, but as is seen in this study, not all pixels can be well utilized, and still several measurements would be acquired per sample. Processing several samples simultaneously would also bring some advantage over using point spectrometers which are only able to record a single spectrum at a time.

To further develop the use of IQ as SOC analyzer, some additional applications would be required. Because of the nature of the data (image), it contains a lot of unnecessary information outside the sample. To ease the extraction of desired data, e.g. some automated material recognition algorithm would be suitable to separate the soil from the image. In a similar working setup with this study it would presumably work well with all other materials in the image being plastic. With material recognition it could also be possible to remove the remaining debris from the sample and preserve only data from actual soil.

6.2.2 Applicability for field

One goal of this study was to evaluate the applicability of the Specim IQ as a portable device for SOC measuring in field conditions. Based on the experiences on the field and the imaging results this still has many challenges. Some challenges in this work could have been avoided with better in advance planning of the field work. Due to lack of time and experience before the field work most of the field methods were developed on the fly and could probably be enhanced from sample preparation to imaging setup as well to get better results.

Optimal case of field application would be the acquisition of data directly from field by imaging undisturbed or semi-undisturbed soil (with minor debris removing and mixing of soil surface) and gaining reliable results from this. This is not practical with IQ for a few reasons:

1) Handheld imaging is not possible in practice and a tripod is basically always needed to avoid trembled images due to the semi-long required exposure time of the scan. Using and carrying around a tripod is usually not practical in field conditions when collecting many samples (depending on the terrain). 2) Lighting conditions should be constant at least during the scan. 3) White reference board is needed in every image and gets dirty very easily. 4) The effect of soil moisture on the recorded spectra (Morgan et al. 2009). 5) Battery and memory card limitations, but this is not very problematic as both can be changed to spare ones and one full battery and memory card can already handle up to 100 measurements (Specim IQ manual.).

These issues could be possibly tackled while working e.g in a flat agricultural field with stable weather conditions and a soil moisture logger but this would require a more complicated model that would be able to take the moisture into account. Moisture content was measured in the field for this work, but the measurements were unreliable and introducing it to the models would have required quite a lot of additional work. The lighting conditions could be improved under direct sunlight with white shade to gain diffuse light and avoid oversaturation of the images.

More functional application could be to build a similar imaging setup to this study close to the wanted research area, as the setup requirements are quite low with one or two halogen lights and a tripod,

6.3 Further research suggestions

As laboratory spectroscopic methods are already quite functional and working, future research should be aimed to agile methods that can be applied with minimum requirement for high level equipment and laboratory conditions, to make SOC analyzes faster and more accessible. This includes improving the at-field measurements and field methods whether using point spectrometers or portable imaging spectrometers. The unpredictable field conditions favor point spectrometers or snapshot hyperspectral devices (Jung et al. 2015) that are not as vulnerable to changing lighting conditions as scanning devices, but there are

still other issues affecting also these measurements such as moisture (Morgan et al. 2009) and soil roughness (Jung et al. 2015).

Models for solely recognizing soil from other materials are also needed to get rid of unwanted objects in the images, these could be utilized in a setup such as used in this work as well as with aerial imagery.

7 Conclusions

The feasibility of Specim IQ for SOC prediction in field and in laboratory conditions was studied as well as the functionality of lasso regression as an alternative multivariate method for PLSR. The data consisted of 191 topsoil samples collected from Taita Hills, Kenya with the same samples measured intact in field and in laboratory after oven drying. Three datasets were produced for the laboratory measurements, one with the average sample spectrums, one with 36 average sub-image spectrums per sample, and one reduced average sub-image dataset with 30 spectrums per sample. Field data was tested only with sample average spectrum.

Results with laboratory data were good and at least regional models working for this dataset were achieved. Using the sub-image datasets improved the R^2 score by 0.05 and reduced the RMSE with 0.1 with PLSR but had no significant effect on lasso regression. Overall results with lasso were however good and its potential as a substitute for PLSR was shown. Important wavelengths found by lasso and PLS were in line further indicating that both models were able to find the essential information for SOC prediction from the datasets.

Results with field data were not good with issues mostly related to the field imaging conditions and moist samples, and the true potentiality of IQ in field conditions was maybe not achieved in this work. No R^2 scores over 0.5 were achieved and not much emphasis was put on improving these models due to the poor quality of the data.

Improvements for the models used in this work could be achieved by introducing ancillary datasets with biophysical information of the sample surroundings such as land cover or

vegetation, but as laboratory spectroscopy methods are already in a fairly good level, future research should emphasize on improving the results on field conditions.

8 Acknowledgements

I would like to thank all the people involved in helping me to complete this thesis.

First of all thanks to all the staff of Taita Research station, especially Mr. Mwadime Mjomba and Mr. Darius Kimuzi who greatly assisted me with the field work. Thanks to Ilja Vuorinne and Dr. Janne Heiskanen for collecting the soil samples from the Teita sisal estate. Sincere thanks to Kristiina Karhu for providing me the license to export the samples to Finland and for providing the premises processing and analysing the samples at the Department of Forest Sciences. Also thank you Mariut Wallner from the Department of Forest Sciences for assisting me with the analysing equipment and technicalities. Finally, thanks to my supervisor Professor Petri Pellikka for proposing this thesis subject, organizing the field work and for supporting and helping me throughout the process. Furthermore many thanks to my second supervisor Mikko Toivonen from the Department of Computer Sciences first for borrowing the Specim IQ for this work but also for assisting with the modelling problems and for e.g. suggesting the use of lasso regression, your help was much appreciated.

9 Bibliography

A. D. Bayer, M. Bachmann, D. Rogge, A. Müller & H. Kaufmann, 2016. Combining Field and Imaging Spectroscopy to Map Soil Organic Carbon in a Semiarid Environment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* **9**(9), pp. 3997-4010.

Aldana-Jague, E., Heckrath, G., Macdonald, A., van Wesemael, B. & Van Oost, K., 2016a. UAS-based soil carbon mapping using VIS-NIR (480–1000 nm) multi-spectral imaging: Potential and limitations. *Geoderma,* **275**, pp. 55-66.

Aldana-Jague, E., Heckrath, G., Macdonald, A., van Wesemael, B. & Van Oost, K., 2016b. UAS-based soil carbon mapping using VIS-NIR (480-1000 nm) multi-spectral imaging: Potential and limitations. *Geoderma,* **275**, pp. 55-66.

Alt, V., Gurova, T., Elkin, O., Klimenko, D., Maximov, L., Pestunov, I., Dubrovskaya, O., Genaev, M., Erst, T. & Genaev, K., 2020. The use of Specim IQ, a hyperspectral camera, for plant analysis. *Vavilov Journal of Genetics and Breeding,* **24**(3), pp. 259-266.

Autio, A., Johansson, T., Motaroki, L., Minoia, P. & Pellikka, P., 2020. Climate change constraints and vulnerabilities of smallholder farmers across varying agroecological zones in South-East Kenya. (submitted),.

Barreto, A., Paulus, S., Varrelmann, M. & Mahlein, A., 2020. Hyperspectral imaging of symptoms induced by Rhizoctonia solani in sugar beet: comparison of input data and different machine learning algorithms. *Journal of Plant Diseases and Protection,* **127**(4), pp. 441-451.

Bartholomeus, H., Schaepman, M., Kooistra, L., Stevens, A., Hoogmoed, W. & Spaargaren, O., 2008. Spectral reflectance based indices for soil organic carbon quantification. *Geoderma,* **145**(1-2), pp. 28-36.

Behmann, J., Acebron, K., Emin, D., Bennertz, S., Matsubara, S., Thomas, S., Bohnenkamp, D., Kuska, M.T., Jussila, J. & Salo, H., 2018. Specim IQ: evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. *Sensors,* **18**(2), pp. 441.

Ben-Dor, E., Chabrillat, S., Demattê, J., Taylor, G., Hill, J., Whiting, M. & Sommer, S., 2009. Using imaging spectroscopy to study soil properties. *Remote Sensing of Environment,* **113**, pp. S38-S55.

Ben-Dor, E., Irons, J. & Epema, G., 1999. Soil reflectance. *Remote sensing for the earth sciences: Manual of remote sensing,* **3**, pp. 111-188.

Ben-Dor, E., Inbar, Y. & Chen, Y., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. pp. 1-15.

Cardinael, R., Chevallier, T., Cambou, A., Beral, C., Barthès, B.G., Dupraz, C., Durand, C., Kouakoua, E. & Chenu, C., 2017. Increased soil organic carbon stocks under agroforestry: A survey of six different sites in France. *Agriculture, Ecosystems & Environment,* **236**, pp. 243-255.

Chatterjee, A., Lal, R., Wielopolski, L., Martin, M.Z. & Ebinger, M., 2009. Evaluation of different soil carbon determination methods. *Critical Reviews in Plant Science,* **28**(3), pp. 164-178.

Derron, M. & Jaboyedoff, M., 2019. Preliminary tests of a portable VNIR hyperspectral camera on rock surfaces. *Geophysical Research Abstracts* 2019.

Doetterl, S., Stevens, A., Van Oost, K. & Van Wesemael, B., 2013. Soil organic carbon assessment at high vertical resolution using closed-tube sampling and Vis-NIR spectroscopy. *Soil Science Society of America Journal,* **77**(4), pp. 1430-1435.

Dyar, M., Carmosino, M., Breves, E., Ozanne, M., Clegg, S. & Wiens, R., 2012. Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples. *Spectrochimica Acta Part B: Atomic Spectroscopy,* **70**, pp. 51-67.

FAO AND ITPS, 2018. *Global Soil Organic Carbon Map (GSOCmap) Technical Report.* Rome: FAO.

Gobrecht, A., Roger, J. & Bellon-Maurel, V., 2014. Major issues of diffuse reflectance NIR spectroscopy in the specific context of soil carbon content estimation: a review. *Advances in Agronomy.* Elsevier, pp. 145-175.

Hall, R., 2008. *Soil essentials: managing your farm's primary asset.* Landlinks Press.

Hbirkou, C., Pätzold, S., Mahlein, A. & Welp, G., 2012. Airborne hyperspectral imaging of spatial soil organic carbon heterogeneity at the field-scale. *Geoderma,* **175**, pp. 21-28.

Heiri, O., Lotter, A.F. & Lemcke, G., 2001. Loss on ignition as a method for estimating organic and carbonate content in sediments: reproducibility and comparability of results. *Journal of Paleolimnology,* **25**(1), pp. 101-110.

Heiskanen, J., Adhikari, H., Piiroinen, R., Packalen, P. & Pellikka, P.K., 2019. Do airborne laser scanning biomass prediction models benefit from Landsat time series, hyperspectral

data or forest classification in tropical mosaic landscapes? *International Journal of Applied Earth Observation and Geoinformation,* **81**, pp. 176-185.

Houghton, R., 1999. The annual net flux of carbon to the atmosphere from changes in land use 1850–1990. *Tellus B,* **51**(2), pp. 298-313.

ICRAF, , Soil-Plant Spectral Diagnostics Lab [Homepage of World Agroforestry], [Online]. Available: https://worldagroforestry.org/sd/landhealth/soil-plant-spectral-diagnostics-laboratory [11/18, 2020].

Jaetzold, R., Schmidt, H., Hornetz, B. & Shisanya, C., 1983. Farm management handbook of Kenya Vol. II. *East Kenya.Ministry of Agriculture, Kenya,* .

Janzen, H., 2004. Carbon cycling in earth systems—a soil science perspective. *Agriculture, Ecosystems & Environment,* **104**(3), pp. 399-417.

Jensen, J.R., 2009. *Remote sensing of the environment: An earth resource perspective 2/e.* Pearson Education India.

Jensen, J.R., 1996. *Introductory digital image processing: a remote sensing perspective.* Prentice-Hall Inc.

Jung, A., Vohland, M. & Thiele-Bruhn, S., 2015. Use of a portable camera for proximal soil sensing with hyperspectral image data. *Remote Sensing,* **7**(9), pp. 11434-11448.

Kruglikov, N., Danilenko, I., Muftakhetdinova, R., Petrova, E. & Grokhovsky, V., 2019. Spectral characteristics of the meteoritic material after the modeling of thermal and shock metamorphism, *AIP Conference Proceedings* 2019, AIP Publishing LLC, pp. 020227.

L3Harris Geospatial, , Push Broom and Whisk Broom Sensors. Available: https://www.l3harrisgeospatial.com/Support/Self-Help-Tools/Help-Articles/Help-Articles-Detail/ArtMID/10220/ArticleID/16262/Push-Broom-and-Whisk-Broom-Sensors [09/24, 2020].

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. *Science (New York, N.Y.),* **304**(5677), pp. 1623-1627.

Lazaridis, D.C., Verbesselt, J. & Robinson, A.P., 2011. Penalized regression techniques for prediction: a case study for predicting tree mortality using remotely sensed vegetation indices. *Canadian Journal of Forest Research,* **41**(1), pp. 24-34.

Martens, H. & Naes, T., 1992. *Multivariate calibration.* John Wiley & Sons.

Morgan, C.L.S., Waiser, T.H., Brown, D.J. & Hallmark, C.T., 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma,* **151**(3-4), pp. 249-256.

Mwalusepo, S., Massawe, E. & Johansson, T.P., 2016. Spatially continuous dataset at local scale of Taita Hills in Kenya and Mount Kilimanjaro in Tanzania. *Data in brief,* .

Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A. & Kent, J., 2000. Biodiversity hotspots for conservation priorities. *Nature,* **403**(6772), pp. 853-858.

Nanni, M.R. & Demattê, J.A.M., 2006. Spectral reflectance methodology in comparison to traditional soil analysis. *Soil Science Society of America Journal,* **70**(2), pp. 393-407.

Nelson, D.W. & Sommers, L.E., 1996. Total carbon, organic carbon, and organic matter. *Methods of soil analysis: Part 3 Chemical methods,* **5**, pp. 961-1010.

Njeru, C.M., Ekesi, S., Mohamed, S., Kinyamario, J., Kiboi, S. & Maeda, E., 2017. Assessing stock and thresholds detection of soil organic carbon and nitrogen along an altitude gradient in an east Africa mountain ecosystem. *Geoderma Regional,* **10**, pp. 29-38.

Njeru, C.M., Ekesi, S., Mohamed, S.A., Kinyamario, J.I., Kiboi, S. & Maeda, E.E., 2017. Assessing stock and thresholds detection of soil organic carbon and nitrogen along an altitude gradient in an east Africa mountain ecosystem. *Geoderma Regional,* **10**, pp. 29-38.

O'rourke, S. & Holden, N., 2011. Optical sensing and chemometric analysis of soil organic carbon–a cost effective alternative to conventional laboratory methods? *Soil Use and Management,* **27**(2), pp. 143-155.

Ontl, T.A. & Schulte, L.A., 2012. Soil carbon storage. *Nature Education Knowledge,* **3**(10),.

Pellikka, P., 1990. Land use and its classi fi cation using a multispectral SPOT XS satellite image in the Taita Hills.

Pellikka, P.K., Clark, B.J., Gosa, A.G., Himberg, N., Hurskainen, P., Maeda, E., Mwang'ombe, J., Omoro, L.M. & Siljander, M., 2013. Agricultural expansion and its consequences in the Taita Hills, Kenya. *Developments in Earth surface processes.* Elsevier, pp. 165-179.

Pellikka, P., Heikinheimo, V., Hietanen, J., Schäfer, E., Siljander, M. & Heiskanen, J., 2018. Impact of land cover change on aboveground carbon stocks in Afromontane landscape in Kenya. *Applied Geography,* **94**, pp. 178-189.

Pellikka, P.K.E., Lötjönen, M., Siljander, M. & Lens, L., 2009. Airborne remote sensing of spatiotemporal change (1955–2004) in indigenous and exotic forest cover in the Taita

Hills, Kenya. *International Journal of Applied Earth Observation and Geoinformation,* **11**(4), pp. 221-232.

Peng, X., Shi, T., Song, A., Chen, Y. & Gao, W., 2014. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. *Remote Sensing,* **6**(4), pp. 2699-2717.

Peón, J., Recondo, C., Fernández, S., F Calleja, J., De Miguel, E. & Carretero, L., 2017. Prediction of topsoil organic carbon using airborne and satellite hyperspectral imagery. *Remote Sensing,* **9**(12), pp. 1211.

Post, W.M., Peng, T., Emanuel, W.R., King, A.W., Dale, V.H. & DeAngelis, D.L., 1990. The global carbon cycle. *American Scientist,* **78**(4), pp. 310-326.

Ramesh, T., Bolan, N.S., Kirkham, M.B., Wijesekara, H., Kanchikerimath, M., Rao, C.S., Sandeep, S., Rinklebe, J., Ok, Y.S. & Choudhury, B.U., 2019. Soil organic carbon dynamics: Impact of land use changes and management practices: A review. *Advances in Agronomy.* Elsevier, pp. 1-107.

Richards, J.A. & Richards, J., 1999. *Remote sensing digital image analysis.* Springer.

Rossel, R.V., Walvoort, D., McBratney, A., Janik, L.J. & Skjemstad, J., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma,* **131**(1-2), pp. 59-75.

Scharlemann, J.P., Tanner, E.V., Hiederer, R. & Kapos, V., 2014. Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Management,* **5**(1), pp. 81-91.

Smith, K.A. & Tabatabai, M.A., 2003. Automated instruments for the determination of total carbon, hydrogen, nitrogen, sulfur, and oxygen. *Soil and Environmental Analysis; Modern Instrumental Techniques,* , pp. 202-246.

Specim IQ manual, , https://www.specim.fi/downloads/iq/manual/software/iq/topics/specim-iq-introduction.html. Available: https://www.specim.fi/downloads/iq/manual/software/iq/topics/specim-iq-introduction.html.

Steffens, M. & Buddenbaum, H., 2013. Laboratory imaging spectroscopy of a stagnic Luvisol profile—High resolution soil characterisation, classification and mapping of elemental concentrations. *Geoderma,* **195**, pp. 122-132.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L. & Van Wesemael, B., 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma,* **158**(1-2), pp. 32-45.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* **58**(1), pp. 267-288.

Todd-Brown, K., Randerson, J., Post, W., Hoffman, F., Tarnocai, C., Schuur, E. & Allison, S., 2013. Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences,* **10**(3), pp. 1717-1736.

Vasques, G.M., Grunwald, S. & Sickman, J.O., 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. pp. 14-25.

Viscarra Rossel, R. & Hicks, W., 2015. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. *European Journal of Soil Science,* **66**(3), pp. 438-450.

Vuorinne, I., 2020. Assessing Agave sisalana biomass from leaf to plantation level using field measurements and multispectral satellite imagery. Helsingin yliopisto, .

Wachiye, S., Merbold, L., Vesala, T., Rinne, J., Rasanen, M., Leitner, S. & Pellikka, P., 2020. Soil greenhouse gas emissions under different land-use types in savanna ecosystems of Kenya. *Biogeosciences,* .

Walkley, A. & Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Science,* **37**(1), pp. 29-38.

Wold, H., 1975. Path models with latent variables: The NIPALS approach. *Quantitative sociology.* Elsevier, pp. 307-357.

Wold, S., Sjöström, M. & Eriksson, L., 2001a. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* **58**(2), pp. 109-130.

Wold, S., Sjöström, M. & Eriksson, L., 2001b. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* **58**(2), pp. 109-130.

Yang, H. & Li, J., 2013. Predictions of soil organic carbon using laboratory-based hyperspectral data in the northern Tianshan mountains, China. *Environmental monitoring and assessment,* **185**(5), pp. 3897-3908.

Zhao, L., Sun, Y., Zhang, X., Yang, X. & Drury, C., 2006. Soil organic carbon in clay and silt sized particles in Chinese mollisols: relationship to the predicted capacity. *Geoderma,* **132**(3-4), pp. 315-323.

10 Appendices

Appendix 1. The python libraries and functions used to optimize the models

Libraries

```
from sys import stdout
import numpy as np
from sklearn.cross_decomposition import PLSRegression
from sklearn.linear_model import LinearRegression, Lasso
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import cross_val_predict
from sklearn.model_selection import train_test_split
```

PLSR optimization

```
def optimise_pls_cv(X, y, n_comp, plot_components=True):
    '''Run PLS including a variable number of components, up to n_comp,
       and calculate MSE '''
    mse = []
    component = np.arange(1, n_comp)
    for i in component:
        pls = PLSRegression(n_components=i)
        # Cross-validation
        y_cv = cross_val_predict(pls, X, y)
        mse.append(mean_squared_error(y, y_cv))
        comp = 100*(i+1)/40
        # Trick to update status on the same line
        stdout.write("\r%d%% completed" % comp)
        stdout.flush()
    stdout.write("\n")
    # Calculate and print the position of minimum in MSE
    msemin = np.argmin(mse)
    print("Suggested number of components: ", msemin+1)
    stdout.write("\n")

    return msemin +1
```

Lasso optimization

```
def optimise_lasso(X,y,niter,plot_components=True):
    """ funktio Lasso regression alpha-arvon määrittämiseksi"""
    mse = []
    alphas = []
    label_alphas = []
    for n in range(niter):
        a = 10**(-0.25*(n+4))
        L = Lasso(alpha=a,max_iter = 1e6)
        y_cv = cross_val_predict(L,X,y,cv=5)
        mse.append(mean_squared_error(y,y_cv))
        alphas.append(a)
        label_alphas.append("%.5f"%a)
        print('alpha: {}, mse: {}'.format(a,mean_squared_error(y,y_cv)))

        if len(mse)>2:
            if (mse[-1] > mse[-2] and mse[-2] > mse[-3]):
                break

    min_mse = np.argmin(mse)
    alpha = alphas[min_mse]


    return alpha
```